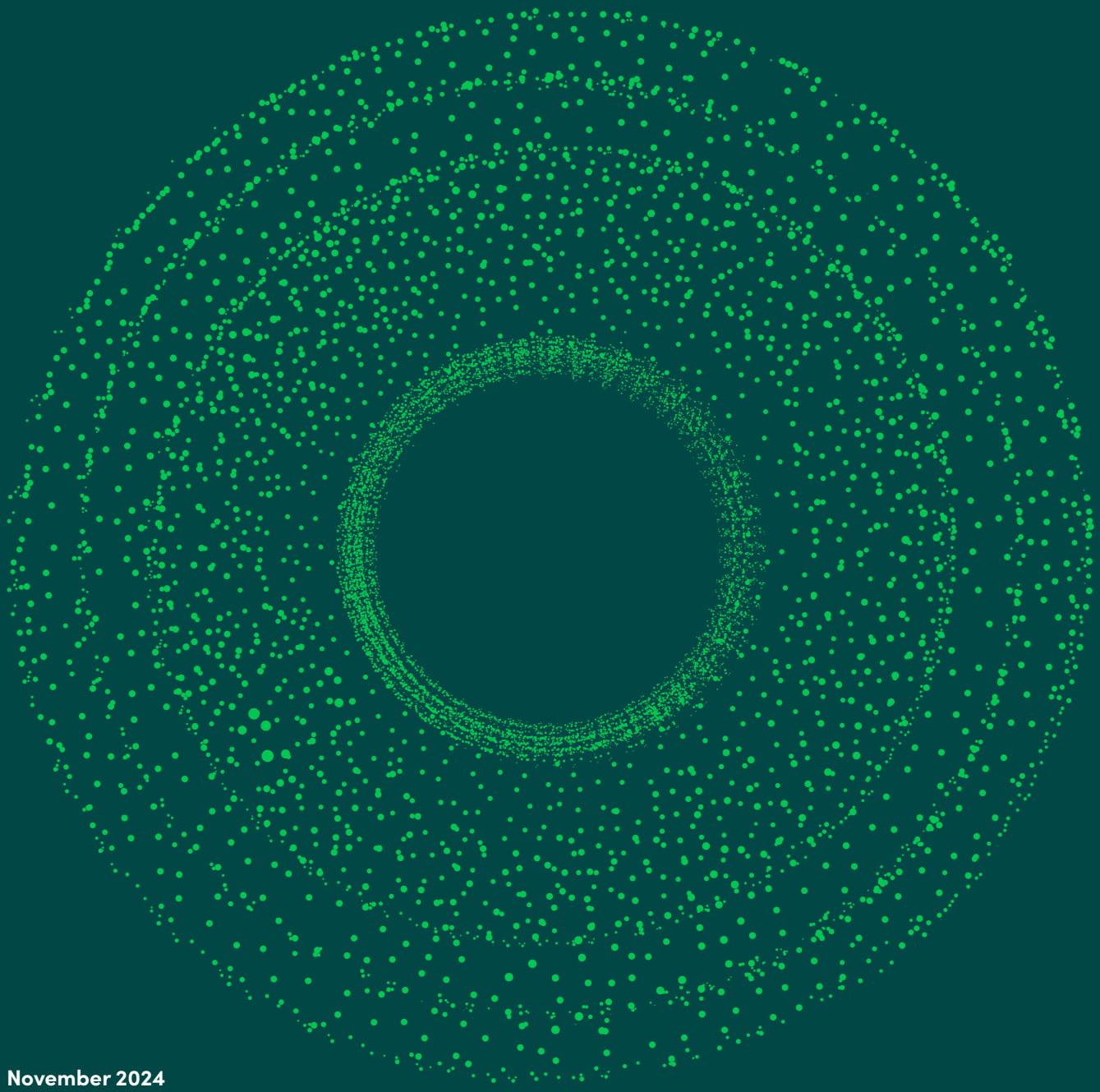
Uniting the UK's Health Data: A Huge Opportunity for Society

A review of the UK health data landscape commissioned by the Chief Medical Officer for England, the UK National Statistician and NHS England's National Director for Transformation

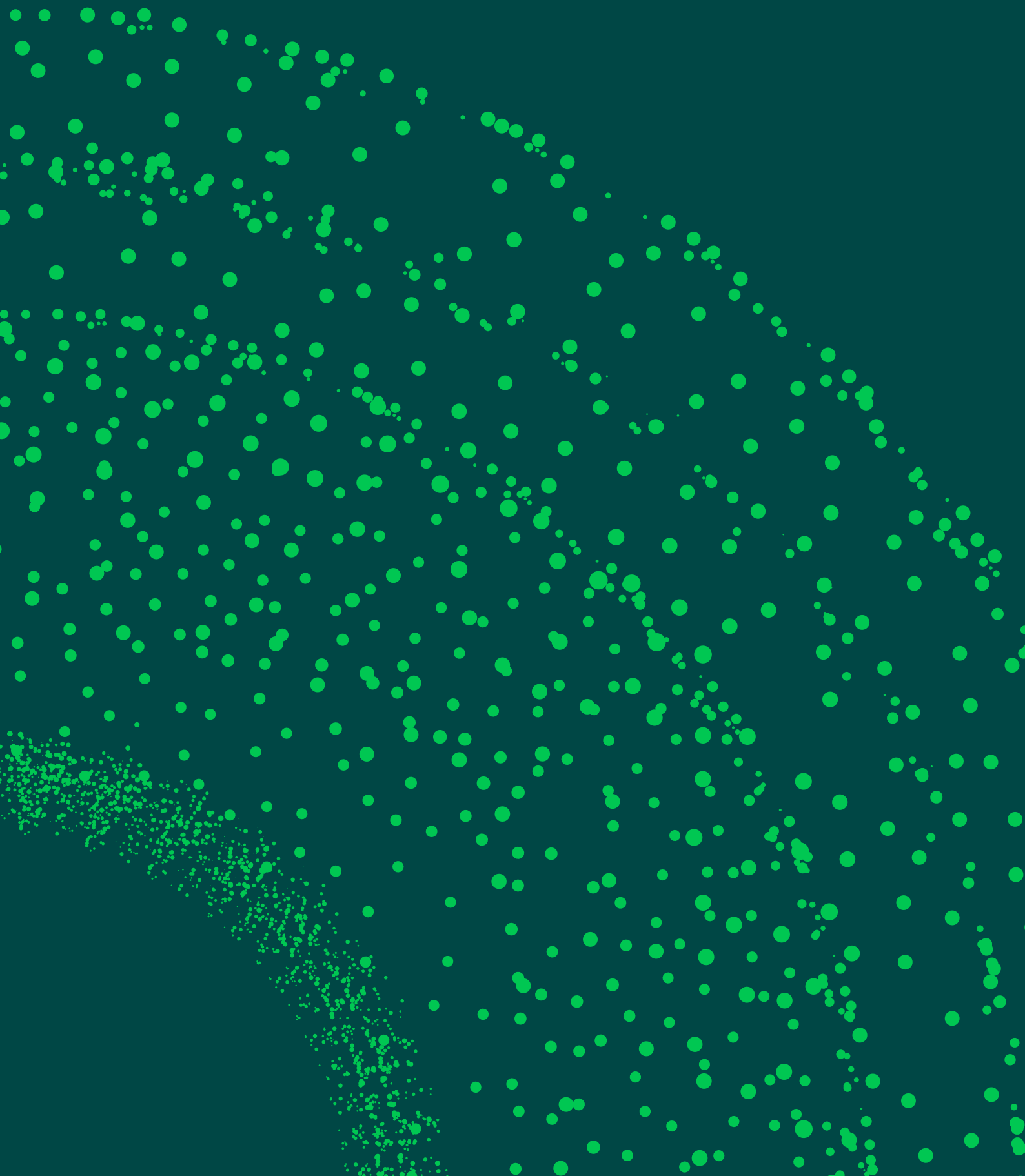


November 2024

A review of the UK health data landscape commissioned by the Chief Medical Officer for England, the UK National Statistician and NHS England's National Director for Transformation

November 2024

Uniting the UK's Health Data
A Huge Opportunity for Society



Contents

Appendices

Foreword

The extraordinary advances in health which have occurred over the last century are driven by science and data. Data can be used in many ways to improve health outcomes for people needing treatment now, and to improve outcomes for citizens and patients in the future.

There are many uses to which data can be put to improve the health of citizens. Data for patient care is the most immediate. Bringing together data from multiple sources can help improve the treatment of current patients, ensuring they get the best diagnosis and most appropriate treatment. Data can also be used operationally to improve the effectiveness and efficiency of the NHS and health and social care systems for all. We can also use data to target effective prevention and public health. Data underpins many areas of research to improve future health, and the scale and speed of analysis now possible will accelerate and strengthen our ability to hand on to our successors much better prevention, diagnosis, treatment and care for disease and disability than was possible even a decade ago. It also allows us to identify people who could benefit from specific clinical trials.

The principle that we should be making the best use of our data is therefore widely supported. The question is how best to do this, and this is less easy. The data we need are widely spread across multiple sources both within the NHS and wider systems and need to be identified and brought together, including non-medical data from other parts of government. Many of the data sources are not designed to be analysed together, even when doing so is clearly in the interests of current and future patients. There are important issues of data safety and privacy that rightly underpin public support.

We therefore commissioned this independent report from Prof Cathie Sudlow, with the support of the relevant Ministers and Chief Medical Officers of the four UK nations. It provides a wide-ranging and very informative review of the UK-wide health data landscape. It goes on to lay out the barriers to safe and trustworthy uses of data for patient and public benefit and makes several very helpful recommendations for overcoming these. We think it is an excellent report that will improve our use of data from multiple sources for the benefit of current and future patients and wider society.

Progress has been made in many areas in recent years, but the case for significantly faster and more systemic change is compelling. We will use Professor Sudlow's findings and recommendations to develop a plan for a national health data research service for England, and to support current and future health across the UK.

Prof Sir Chris Whitty
Chief Medical Officer for England

Prof Sir Ian Diamond
UK National Statistician

Vin Diwakar
National Director of Transformation (interim), NHS England

Dr Tim Ferris
Former National Director of Transformation, NHS England

Personal note from the author

Undertaking this review has been both a daunting task and a great privilege. It has been hugely rewarding to discuss the UK-wide health data landscape with hundreds of people who have generously provided their time, expertise, knowledge, experience, views and advice. I am very grateful to each of them.

Through my various roles, I have been embedded in the UK's health data landscape for many years and I thought I knew it well. But conducting this review has reinforced just how extraordinarily complex this landscape is – one where finding potential solutions to the toughest challenges requires a broad understanding not only of data, science and technology, but also of the health and care system, of government, and of ethical, legal, social, cultural, behavioural, financial, geographical and political factors. Further, the sheer volume of information about the wide range of health-relevant data sources can sometimes seem like a bottomless pit. This is compounded by variation between the UK's four nations as well as by frequently changing organisational labels, structures, strategies and policies. Readers with a deep knowledge of any part of the landscape will inevitably find gaps in this review and for those I apologise.

In my roles as a doctor, and as both a user and creator of large-scale health data resources for research, I have seen the huge benefits for patients and society from a wide range of uses of health-relevant data. But I have also experienced, repeatedly over many years, the frustration of knowing that there are vast amounts of such data that could and should be – but are not – accessed and used to improve patient care, and to advance health research, care and policy for patient and public benefit. Consulting widely in preparing to write this review has confirmed that I am not alone: this frustration is widespread across the NHS, our academic institutes and universities, charities, organisations representing patients and the wider public, the life sciences industry, research funders and government.

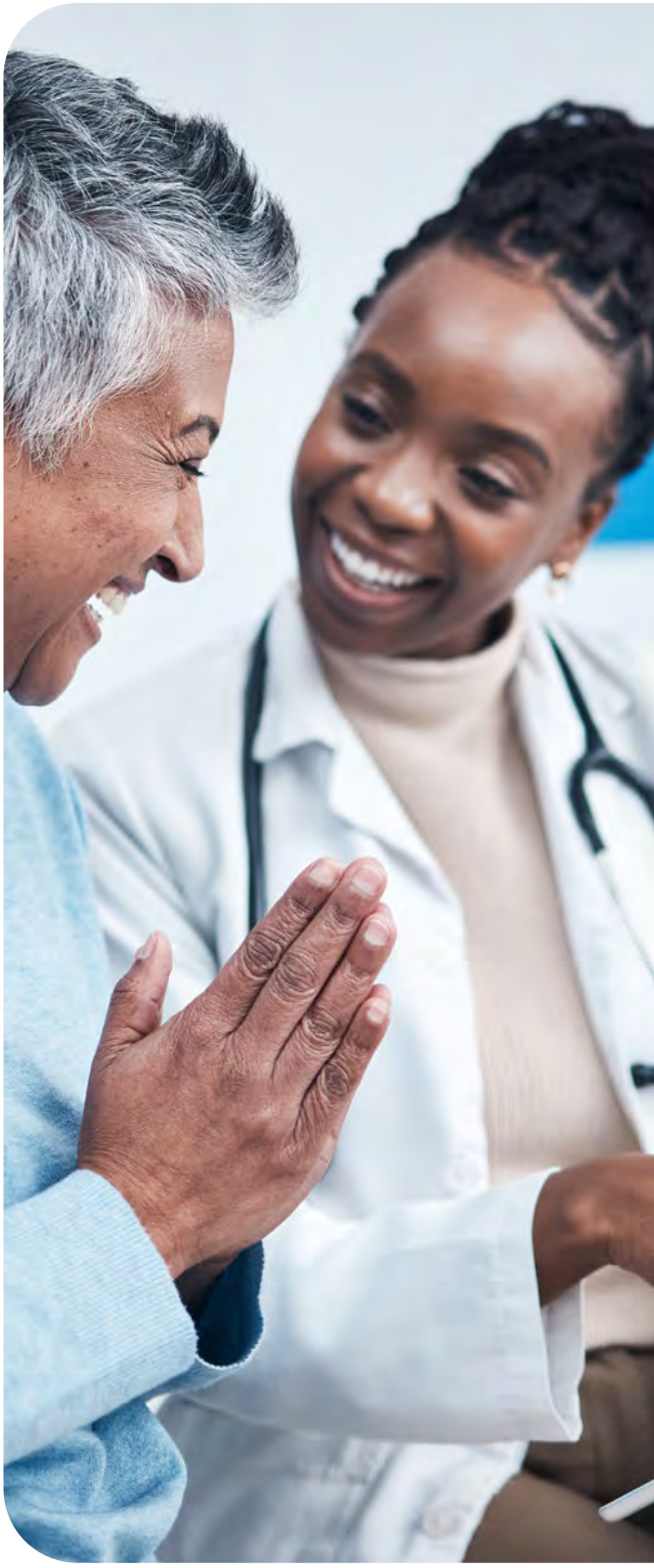
Across the UK, despite key advances in recent years, we are simply not maximising the benefits to society from the many already-existing sources of health-relevant data. In some respects, we are even slipping backwards from some of the excellent progress made during the COVID-19 pandemic in broadening safe, secure access to and use of health-relevant data for patient and public benefit. Far too many lives are unnecessarily lost or ruined because of blockers or delays in safe and secure access to, linkage and analysis of existing health data. These blockers impede the generation of insights to guide and improve our health and care system. They delay or prevent hundreds of medical and population health research studies that collectively involve millions of people across the UK. These studies are essential to improving our health and wellbeing, through their aims to unravel the underlying causes of diseases; to develop better diagnosis, prevention and treatment strategies; to test these in clinical trials; and to undertake analyses in whole populations to assess their uptake, effectiveness and safety in the 'real world'.

There is a huge opportunity to turn this backward slippage around, to capitalise on the UK's substantial health data assets, and to accelerate towards a future where the power of health-relevant data is fully realised. This will only happen if we work collectively across the UK to simplify the overly complex landscape and to lay out a coherent vision and roadmap, where benefit for patients and the wider public is the key motivating goal.

Professor Cathie Sudlow, OBE

Key points

- Every day, health and care professionals, researchers and policymakers use health data safely to improve people's health and lives.
- People in the UK overwhelmingly support the use of their health data, with appropriate safeguards, to benefit themselves and others.
- The UK has abundant sources of data relevant to our health, both from its unique National Health Service and a range of other sources.
- The most powerful insights come from linking the different sources of data together.
- But health and care professionals, researchers and policymakers face many obstacles and delays in accessing, linking and analysing health data to improve people's care and lives.
- These barriers arise from the UK's complex and inefficient systems for managing and accessing health data.
- They prevent health and care professionals from accessing all the information they need to provide the best patient care.
- And they prevent or delay crucial analysis and research about health conditions affecting millions of people across the UK.
- We are letting patients and their families down as a result.
- We need to recognise our national health data for what they are: critical national infrastructure that can underpin the health of the nation. They should be treated as such with a strategy, leadership and investment to match.
- This review sets out how that can be achieved with five key recommendations.



Executive summary

All aspects of our lives are increasingly captured in digital form, and the health and care system is no exception.

An abundance of data is generated each second of every day through our many encounters with the National Health Service (NHS) when we visit our general practitioner (GP), spend time in hospital or have a blood test or scan. There are also data relevant to our health from social care, education, justice, earnings and disability settings, not forgetting data from those that take part in population or clinical research studies and constant monitoring of the weather and pollution levels, all while our phones and devices measure step count, heart rate and sleep patterns.

We have a huge opportunity, and indeed a responsibility, to use health data safely and securely to improve health, wellbeing and prosperity across society.

Health data can be accessed for uses beyond our own direct care, for example when people give informed consent for use of their data, or when approved studies for public benefit access large datasets that have been stripped of patient-identifying information and held in protected environments.

Lord Darzi's recent review of the state of the NHS in England[1] highlights the critical condition of our health and care system. He calls for a major 'tilt towards technology' as one of the ways to improve the NHS, and points to the potential of AI and life sciences breakthroughs to transform care and treatments. These advances will rely on the effective and trustworthy use of health data.

As **Chapter 1** of this review sets out, health data can and should be used to:

- support the care of each one of us if we become sick;
- inform an intelligent health and care system capable of predicting and responding to varying demands, for example managing outpatient and operating theatre waiting lists and addressing bottlenecks in hospital emergency departments and in discharges from hospital;
- support the planning and equitable delivery of health, care and public health services that meet the needs of local, regional and national populations, keeping people healthy through preventing ill health as well as looking after them when they are unwell;
- enable a wide range of research and innovation to benefit patients and the public.

1 See https://www.gov.uk/government/publications/independent-investigation-of-the-nhs-in-england.

Improving and saving lives

The four nations of the UK have a long history of collecting national databases of health-relevant data from their entire populations, going back as far as the 1950s in Scotland. Almost all the 67 million people living in the UK receive most of their healthcare from the NHS. This makes our national collections of data amongst the largest and most comprehensive worldwide.

As a result, the UK hosts some of the best examples globally of transformational use of health data for public benefit.

- UK Biobank is a large-scale database and research resource used by over 30,000 researchers worldwide to better understand the causes and consequences of many different health conditions, such as heart disease, stroke, cancer, diabetes, arthritis, mental health conditions, dementia and many others. It is also used to develop new approaches for their prevention, treatment and diagnosis. More than 500,000 volunteers have undertaken extensive questionnaires, measures and imaging, donated samples and given permission for their health to be followed through their routinely collected NHS and other health-related records. This means UK Biobank has unparalleled depth and breadth of data and samples for carrying out high-quality research.
- Building on the success of UK Biobank, Our Future Health aims to be the UK's largest ever health research programme, with a target of five million volunteers. The programme is designed to support multiple research initiatives to discover and test more effective approaches to prevent, detect and treat diseases. It has partnered with national NHS organisations such as NHS England, using centralised NHS databases to issue invitations to take part to millions of eligible people across the UK. As a result, well over one million volunteers have already joined the programme. Their health and wellbeing will be followed over many years through linking to national health-related records from the NHS and many other sources.
- During the pandemic, UK policymakers were rapidly informed about the impact of COVID-19 infection and vaccination on people with different health conditions and of different ages, ethnicities and socio-economic circumstances. This was made possible through the secure linkage and analysis of a variety of health datasets for the whole populations of the four UK nations.

However, these major UK-led successes in health data-driven research are far too often the exception rather than the rule. The complexity of our data systems and data governance means such work is far from routine when it could be business as usual.

The national response to the COVID-19 pandemic drove remarkable progress in broadening secure access to health-relevant data for patient and public benefit. The data-driven RECOVERY trial is a fantastic example of this. It was able to answer key questions about how to treat severe COVID-19 and led to the widespread use of treatments such as dexamethasone, saving hundreds of thousands of lives worldwide. Yet some of these advances in the use of data to inform healthcare are now falling back to pre-pandemic approaches.

Existing national data collections should be relatively straightforward for approved researchers to access, link and analyse. But in practice, access is difficult, slow or impossible. For example, access to data currently available via NHS England often takes many months or even years.

Furthermore, many national NHS England data are only accessible for COVID-related analysis and research (for example general practice data and national cardiovascular audits) but not to tackle other health conditions, such as other infectious diseases, cancer, heart disease, stroke, diabetes and dementia.

And more complex types of health data generally do not have national data systems (for example, most laboratory testing data and radiology imaging). Existing examples of what can be done (for example Scotland's national medical imaging database) make national solutions for these complex data ambitious but achievable goals.

The existing barriers impede work to guide and improve our health and care system. They delay or prevent hundreds of health research studies that could:

- improve our health and wellbeing by unravelling the causes of diseases;
- develop better diagnosis, prevention and treatment strategies for conditions affecting many millions of people;
- test these in clinical trials and assess their uptake, effectiveness and safety in the 'real world'.

Support from patients, public and professionals

People in the UK overwhelmingly support the use of their health data to benefit themselves and others. Surveys, in-depth focus groups and other information-gathering exercises over the last 15–20 years have consistently shown this, as summarised in **Chapter 2**.

Most people want to know how their data are being and will be used. They are concerned to know how their privacy will be protected and their data kept secure; that robust and transparent mechanisms are in place to ensure that data are used for public good; and that they can choose not to have their data used for certain purposes beyond their direct clinical care.

Many people want to be able to access and, where needed, amend their own health records. And many are more cautious about uses of data from which organisations might profit financially.

Less is known about the views of healthcare professionals. General practice data are some of the most important for improving healthcare. But, while some GPs are at the forefront of efforts to ensure data are used widely for patient and public benefit, we know that others are more circumspect than patients and members of the public. Some GPs have concerns about inadvertently breaching laws that protect the confidentiality and privacy of patient data. Some may also be concerned about data being used for performance management. However, despite these concerns, GPs are generally supportive of wider uses of health data for patient and public benefit. It is important that measures to increase the use of data reduce, or do not add to, the burdens of general practice.

The complex health data ecosystem: data from many sources, not just the health service

Fulfilling the potential of health data to improve lives is not straightforward. It can be beset by delays within a system that is frustrating to navigate. This is caused by the complexity and fragmentation of the health data ecosystem.

To start with, the NHS is not one organisation but many – general practices, hospital and mental health trusts, integrated care boards, and more – and its constituent parts do not always work together effectively.

The digitisation of healthcare in the UK lags behind other high-income countries and other sectors. Indeed, parts of the health and care system, social care in particular, still depend on paper records. The piecemeal introduction of many different computer systems into the health and care system, provided by many different companies, has created difficulties in 'interoperability' – the ability of these different systems to talk to each other. People working in multiple NHS bodies spend a lot of time maintaining the many contracts needed between their organisations and multiple commercial computer system suppliers.

Added complexity comes from the statutory and common law frameworks that govern access to, and uses of, health-relevant data. These are complicated to start with and are interpreted and applied differently by the many data custodian organisations across the complex ecosystem. Furthermore, the common law position and mechanisms for complying with it differ between the four nations of the UK.

Amid this complexity, this review was commissioned to map the health data landscape across the UK. **Chapter 3** is by far the longest in the review and provides a guide to the many sources of health-relevant data: the health data collected and held by general practices, hospitals, laboratories, X-ray and scanning departments, high street opticians, pharmacies, dentists and others, as well as the health-relevant data coming from many sources beyond the NHS. We hope this chapter will prove useful to all those using data to improve healthcare, whatever their perspectives on how it should be accessed.

Each of the various sources of health-relevant data can separately provide useful information. But their power comes from when datasets are linked together. This is when the most important and transformational insights emerge, as **Chapter 4** demonstrates.

For example, we can only really know if breast cancer screening is improving cancer outcomes by connecting the data from national breast screening programmes to data on cancer cases and long-term survival from national cancer registration systems. And we can only fully understand the impact of ill health on employment status and economic activity by connecting data from health records to data on earnings. Such understanding is crucial to inform policies to improve the current situation, one where economic inactivity due to long-term sickness in the UK has reached a record high of 2.8 million people, representing a key risk to the economy, the government's fiscal position and the NHS.

However, far from being routine, successful linkages of different data sources are all too uncommon, especially those that bring together data from the NHS with data from other sectors, such as census, education or earnings. In England, barriers to such cross-sectoral linkages include the lack of streamlined data sharing processes between NHS England and the Office for National Statistics (ONS), and lack of clarity about how to comply with the common law duty of confidentiality when linking health and non-health data.

Safe and secure data access

Everyone's health records contain sensitive information that is personal and private. There is a human story behind each data point, and the privacy, confidentiality and security of health data must be taken extremely seriously by all.

The internationally accepted 'Five Safes Framework' (**safe data, safe research, safe people, safe settings, safe outputs**) is widely used to guide research and analysis using health data. Designed by UK experts, the framework protects the privacy and security of people's data, ensures that data are used for the public good, and guards against misuse.

To comply with this framework, data custodian organisations:

- de-identify (remove any information such as NHS number that could directly identify a patient) or completely anonymise data (**safe data**) wherever possible;
- make data available only for approved uses for public benefit (**safe research**);
- make data available only to appropriately trained, certified and authenticated analysts (**safe people**);
- provide data wherever possible within highly secure computing environments called secure data environments (SDEs) (**safe settings**). SDEs operate like a reading library rather than a lending library, in that analysts cannot download or export any person-level data and must leave them where they are;
- check the results of any analysis (for example tables or figures) before they are exported from SDEs to ensure that they could not be used to identify any individual (**safe outputs**).

Recent years have seen a network of SDEs develop across the four nations of the UK, providing analysts with secure remote access to de-identified health data in a protected environment.

Chapter 5 explains how SDEs across the four nations enable access to national-level NHS datasets, data at regional levels and data from other sectors outside the NHS.

Barriers to using health data for public benefit

Several barriers need to be overcome to enable and encourage more beneficial uses of health data. These barriers and potential solutions are outlined in **Chapter 6**.

- Long-term investment in national health data infrastructure is needed, rather than short-term initiatives with unrealistic timelines for delivery.
- Streamlined processes, economies of scale and reducing unnecessary complexity are essential to make the most of limited resources.
- Data custodian and controller organisations should be rewarded for rapid, efficient and secure access to data and services that improve the productivity of data users in generating public benefit, while maintaining the security of the data.
- Strategic partnerships between the health and care system, government bodies, academic research institutions, major public and charitable funding agencies and the life sciences industry are needed to fill the substantial capacity gaps in information governance and in data management and curation, especially within NHS England.

Recommendations

We make five recommendations for transforming the national health data ecosystem and overcoming these barriers. Although these focus on England, reflecting what the commissioners of this review asked for, their principles apply across all four UK nations. The recommendations are explained in full in **Chapter 7**.

Above all, we need to recognise our national health data for what they are: critical national infrastructure that can underpin the health of the nation. They should be treated as such with a strategy, leadership and investment to match.

We also recommend the establishment of a national health data service for England, embedded within existing organisational structures but with accountable senior leadership and a ring-fenced budget. Its main purpose would be to oversee a service to support streamlined, secure research and analysis of health data by approved analysts. It would establish a single data access system for datasets held across England and would follow a priority list for enabling access to key NHS data assets, starting with general practice data. It would work with the ONS, the UK Statistics Authority and relevant organisations in the UK's devolved nations to develop, improve and streamline mechanisms for the sharing and linkage of data across sectors (specifically linkage of NHS health data to health-relevant data from other settings) and across UK nations. We note that there may be similarities or overlap in this recommendation with the government's emerging plans for a National Data Library.

None of this can be achieved without the ongoing support of patients, the public and health professionals. Ongoing engagement with and meaningful involvement of these groups is strongly recommended to shape these advances, ensure transparency on how health data is being used and inform a single opt-out system.

In summary, the UK's complex and inefficient data systems prevent and delay crucial analysis of health conditions affecting millions of people across the UK. We are letting patients and their families down and no change is not an option.

We must focus on **making the simple easy** (for example access to national datasets that already exist) **and the difficult possible** (for example linking national NHS data to datasets from beyond the NHS). This will require coordinated action across multiple organisations and stakeholders to ensure the greatest benefits for patients and the public from health-relevant data across the UK.

Getting this right holds a great prize. Efficient, effective and trustworthy access to our rich abundance of health data will lead to a step change in the UK's research and innovation capability, enhance healthcare, health service planning and delivery, and bring significant economic and societal benefit to the whole country.

Key recommendations

Recommendation One

Major national public bodies with responsibility for or interest in health data should agree a coordinated joint strategy to make England's health data a critical national infrastructure

Making health data a critical national infrastructure will boost analysis and research to improve health, wellbeing and economic productivity. We recommend that all major national public bodies that generate, collect, manage, curate, fund or use health-related data in England should sign a commitment to: reduce ecosystem complexity; coordinate long-term planning and investment in publicly funded health data infrastructure; support a national health data service; ongoing nationally coordinated engagement with patients, public, health professionals and politicians; a UK-wide strategy for data access and trustworthy governance; and a UK-wide system for SDE standards and accreditation.

Recommendation Two

Leading government health and research bodies should establish a national health data service for England with accountable senior leadership

A national health data service will accelerate research and analysis that benefits society. NHS England (NHSE), National Institute for Health and Care Research (NIHR), the Departments of Health and Social Care (DHSC) and Science, Innovation and Technology (DSIT), and UK Research and Innovation (UKRI) should establish this service to support research and analysis using health data, delivered in partnership with academic, NHS and industry-based research users. It should be led by a senior executive director and have a ring-fenced budget and regularly published performance metrics. Its core functions would be to:

- establish and oversee a **single national health data access system** for England;
- lay out a clear roadmap for data services and dataset provision, complementary national and regional data infrastructures, and streamlined, standardised data governance and access;
- work with the devolved nations, the Office for National Statistics and the UK Statistics Authority to deliver secure, efficient, cross-UK and cross-sectoral data sharing, access and linkage;
- implement an acceptable, transparent investment strategy for health data infrastructure and models for data access cost recovery and pricing.

Key data priorities for this national service should be to:

- **establish a national system for general practice data**, enabling secure access to comprehensive, whole-population, structured, coded general practice data, linkable to other data sources and accessible for a wide range of beneficial uses;
- **enhance and accelerate access to other major national and regional NHS data assets**: hospital episodes, medicines data, lab data (including genomics), national audits and registries, screening data and unstructured clinical data (including imaging and free text);
- **transform access to data from other sectors linked to health and care data at national scale.**

Recommendation Three

The Department of Health and Social Care should oversee and commission a strategy for ongoing coordinated engagement with patients, public, health professionals, policymakers and politicians

The DHSC should commission a coordinated, multi-organisational strategy for ongoing engagement across society. The wide range of potential data uses should be shaped by the input and experience of patients, public and health professionals, while understanding how best to provide transparency of how data are used for all groups. Major areas of emphasis should be better understanding the perspectives of health professionals, especially GPs; accelerating patients' access to their own health data; and informing a single, centralised, national health and care data opt-out system in England, without imposing any burden on busy GPs.

Recommendation Four

The health and social care departments in the four UK nations should set a UK-wide approach for data access processes and proportionate data governance

A UK-wide approach to streamline data access processes and foster proportionate and trustworthy data governance will enable more and better health data analysis and research. The aim should be for trusted researchers and analysts conducting responsible analyses in the public good to be able to rapidly access the de-identified data they need, while ensuring that data cannot be inappropriately accessed. The approach should be set by the health and social care departments of the UK's four nations and developed with patient and public involvement. It should confront legal and regulatory complexity by providing clear guidance on current approaches, proposing improvements that reduce unwarranted variation, and recommending where new or revised legislation is needed.

Recommendation Five

National organisations in the four UK nations should develop a UK-wide system for standards and accreditation of SDEs holding data from the health and care system

The increasing use of SDEs for maintaining greater control over the sensitive health data accessed by approved researchers has been a great advance in recent years. With more SDEs being set up all the time, a UK-wide system for standards and accreditation of SDEs will accelerate the safe use of health data for patient and public benefit. The UK Statistics Authority and health and social care departments in the four UK nations, with input from relevant UK-wide organisations such as Health Data Research UK, Administrative Data Research UK and Data and Analytics Research Environments UK, should lead on the development of: a UK-wide accreditation scheme for SDEs holding data from the health and care system; UK-wide SDE standards to improve user experience and to promote positive user behaviours that benefit all users; and UK-wide policy on avoiding an excess of SDEs.

Introduction

Aims of this review

This review was commissioned in late March 2023 by Professor Sir Chris Whitty (Chief Medical Officer for England and Chief Medical Adviser to the UK Government), Professor Sir Ian Diamond (UK National Statistician and principal adviser on official statistics to the UK Statistics Authority and UK Government) and Dr Tim Ferris (National Director of Transformation for NHS England from May 2021 to September 2023), supported by – amongst others – the Chief Medical Officers from the other three nations of the UK, the Secretary of State for Health and Social Care and the Chief Executive of NHS England. The original request and terms of reference for the review are provided in Appendix 2.

They requested two key tasks:

1. Map the linkable health-relevant datasets across the UK;
2. Outline barriers to sharing data for public benefit, whilst keeping it secure, and identify solutions to overcome these (with a focus on England for this second task).

Other relevant reviews

This review follows in the wake of many other policy documents, reports and reviews from the last few years. These include: Professor Ben Goldacre's review about better, broader and safer approaches to using England's health data for research and analysis;[2] Lord James O'Shaugnessy's review of commercial clinical trials in the UK;[3] and the Tony Blair Institute's recent report about harnessing data for health.[4] Appendix 3 provides further details and background on some of the most relevant recent reviews. Key areas of coverage are summarised in Box A.

2 See https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis.

3 See https://www.gov.uk/government/publications/commercial-clinical-trials-in-the-uk-the-lord-oshaughnessy-review.

4 See https://www.institute.global/insights/politics-and-governance/a-new-national-purpose-harnessing-data-for-health.

Box A Summary of other relevant reviews

- Several previous reviews have made recommendations about how the health and care system must embrace the challenges and opportunities of the data and digital revolution. Governments in each of the four nations of the UK have repeatedly committed to investing in and developing an increasingly digitised NHS. Their aim is for digital capability and insights from data to drive improvements in patient care, planning and delivery of healthcare, social care, public health measures and health research.
- Previous documents on government strategies for the life sciences and clinical research have highlighted major UK research successes during the COVID-19 pandemic. They emphasise that the approaches underpinning these, many of which were dependent on and driven by health data, must be continued and enhanced. They also highlight important gaps to be plugged by building and maintaining essential infrastructure and partnerships across our universities and research institutes, the NHS and social care system, public and charitable research funders, life sciences companies, the Medicines and Healthcare products Regulatory Agency (MHRA), health technology appraisal bodies and – crucially – patients and the public.
- Data relevant to our health are also generated beyond the health and care system (for example education, income and census data). Hence, the UK government's National Data Strategy and reviews highlighting the public benefits of sharing and linkage of data across sectors and government departments, together with proposals to address ongoing challenges, are also pertinent.
- Past reviews highlight the need for ongoing, meaningful deliberation and engagement with patients and the public. They emphasise the importance of transparency and clarity on:

i. the substantial benefits as well as the risks of using health-relevant data for public benefit; and

ii. how, for what purposes, and by whom health-relevant data are (or might be) used.

- Many previous reviews focus mainly or only on one UK country. Some concentrate on data from the health and care system. Others take a much broader perspective on uses of data for public good from multiple sectors, including – but not focusing specifically on – data relevant to health.

Distinctive approach of this review

This review takes a UK-wide approach in mapping health-relevant data across the four nations. While recognising the central importance of data from the health and care system, we consider the breadth of data relevant to health, including data generated beyond this system. We provide background on the various sources and types of health-relevant data. We discuss how these have been, are being, are not being, or could be linked, accessed and used to generate insights for patient and public benefit. We also discuss ongoing publicly funded initiatives that aim to broaden safe and secure access to health-relevant data for a range of beneficial uses.

In addressing barriers and solutions to wider data access and use, we highlight several key priority areas for action. We subdivide these into:

i. processes and systems used to enable access to data; and

ii. sources and types of data of particularly high value, for which overcoming barriers to access would bring substantial patient and public benefit.

Finally, we make several ambitious but hopefully practical recommendations. These aim to maximise the benefits to patients and the wider public, whilst taking a robust yet proportionate approach to managing data privacy and security risks. They consider the need to identify, understand and reduce unnecessary complexity, duplication of effort and costs to the public purse (i.e. the taxpayer), as well as to maintain and comply with ethical, legal and regulatory requirements and standards.

Review consultation methods

From April 2023, the lead author and a small support team engaged extensively with a wide range of relevant stakeholders, gathering information, views and insights to inform this review. Our methods are summarised in Box B.

We focused throughout on asking questions, listening carefully to, and distilling the knowledge, views, expertise and experience of all the stakeholders who engaged with us. Our own background knowledge, experience and expertise in research and healthcare have – of course – informed our approach and thinking. These have been influenced by the extensive consultations undertaken for this review but also – inevitably – by formal and informal interactions with patients and colleagues over many years, as well as by our own personal lives and careers.

We were delighted by the huge interest, enthusiasm and commitment shown by so many in engaging with us. Importantly, although there were mixed opinions and debates around how best to tackle some barriers, every person or organisation who articulated a view to us felt that health-relevant data can and should be used more widely to benefit the health, care and wellbeing of patients and the wider community.

Appendix 4 lists the people (and their organisations) who contributed to one or more discussions. Appendix 5 summarises the findings from the online survey and public facing workshops that formed part of the evidence informing this review.

Box B Summary of review consultation methods

- A series of over 100 semi-structured discussions with individuals or groups representing a broad range of relevant organisations from across all four nations of the UK, including the NHS, patient and public-facing organisations, government, universities and research institutes, life sciences industries, charities, privacy groups. We used these discussions to gather relevant information for the review and to identify additional stakeholders to follow up with. We consulted several times with some individuals and groups to gain a deeper understanding of key barriers, potential solutions and latest updates on relevant ongoing initiatives.
- An open, online survey during spring and summer 2023, which attracted around 180 semi-structured, written responses from a similarly broad range of stakeholder individuals and organisations.
- Two public-facing online workshops (held in August and September 2023), attended and contributed to by around 100 members of the public.
- Meetings convened by relevant national organisations, to discuss certain issues in more depth, or to share and obtain feedback on emerging priorities.
- Intermittent discussions with the commissioners of this review and several of their colleagues, to ensure we were addressing their brief in as helpful, constructive and balanced a way as possible.

We structured our discussions, survey and workshops around three main topics:

1. The data types and flows that should be prioritised and enabled to support not only the delivery of care to individual patients but also a broad range of uses for wider patient and public benefit;
2. Barriers to enabling flows, linkages between and access to these priority data;
3. Solutions to overcoming these barriers and better realising the benefits.

Chapter 1

Using health-relevant data for patient and public benefit: the opportunity

In this chapter

Huge value to society can be gained from researchers and policymakers accessing, analysing and deriving insights from multiple sources of data relevant to our health. This is not something fundamentally new; it has been the case for many decades. However, the potential pace and scale of the opportunity has changed substantially over the last 30 years. A rapidly accelerating transition from paper- to computer-based recording and handling of information has occurred – or is occurring – across multiple sectors, including health and care. We have seen major advances in data storage capacity, data security mechanisms, computing power and data analysis methods, including recent, dizzyingly fast-paced developments in artificial intelligence (AI) and its applications. High-throughput laboratory technologies now exist that can rapidly analyse samples (such as blood, saliva or tissue biopsies) from thousands or even millions of people to generate billions of data points about our genes, proteins, metabolic pathways and other detailed molecular data. In the health and life sciences sectors, these advances bring opportunities to better understand, predict, prevent and treat disease, along with the potential for more efficient, effective and equitable delivery of healthcare.

The health and care sector in the UK lags behind other sectors and the health and care sector in many other high-income countries in fully embracing and implementing the digital and data revolution. Most general practices have been computerised for decades and, following developments in the last 5–10 years,[5] around 90% of hospitals across the UK now have an electronic patient record system. But many component parts of the NHS, and the system as a whole, are still far from being fully 'digitally mature'[6,7,8]. Social care providers have even further to go in their progress towards adopting digital systems and achieving digital maturity: in 2022, less than half of social care providers had any form of digital care record, although this situation is improving rapidly.[9]

Reaching and maintaining digital maturity across the health and care system in all four nations of the UK will continue to require several ingredients: long-term investment; consistent, visionary and highly collaborative leadership; ongoing cultural change; and appropriate training for staff, patients and carers. This is not only essential for the efficient and effective delivery of healthcare; it is also necessary if we are to fully realise the broader benefits of the data generated across the health and care system.

5 E.g., following the recommendations of the Wachter Review of health information technology in England https://assets.publishing.service.gov.uk/media/5a8091afe5274a2e87dba8f2/Wachter_Review_Accessible.pdf.

6 An organisation is considered digitally mature if its procedures, processes and methods rely on and adapt to digital information and tools rather than manual resources and paper records. Obtaining an objective assessment of progress over the last decade or more in digital maturity in the health and care sector is difficult due to a lack of comparable digital maturity data in the public domain. However, for some of the more detailed information that is publicly available, see: https://digitalhealthintelligence.net/digital-maturity-acute-nhs-snapshot-report/, an assessment in 2022 of the digital maturity of 132 NHS England acute trusts, using internationally recognised benchmark criteria; and https://www.digihealthcare.scot/digital-maturity-assessment-2023/, an assessment of digital maturity across the majority of Scotland's healthcare and social care landscape in mid-2023.

7 House of Commons Health and Social Care Committee: Digital Transformation in the NHS (2023): https://committees.parliament.uk/publications/40637/documents/198145/default/.

8 Results from a Royal College of General Practitioners (RCGP) survey of GPs in 2023 reported inadequate broadband (38%), PC or laptop software (46%) and ability of IT systems to exchange information with secondary care (65%). See RCGP *Fit for the Future: Reshaping general practice infrastructure in England,* Chapter 3 *Digital Infrastructure* (2023): https://www.rcgp.org.uk/getmedia/2aa7365f-ef3e-4262-aabc-6e73bcd2656f/infrastructure-report-may-2023.pdf.

9 https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data.

Used well, these data can and should:

1. **support the care of individual patients** – by ensuring that all the right information is available at the right time for patients, their carers, and health and care professionals to inform decisions and plans for each patient's care;
2. **inform the operation of an intelligent system** that can work locally, regionally and nationally to provide efficient and effective patient care – for example, to manage appointment booking systems flexibly or to ensure the best use of available hospital ward, operating theatre and intensive care capacity;
3. **support the planning and delivery of health, care and public health services** (for example vaccination or screening programmes) at the level of national (whole-country) or regional populations;
4. **support a wide range of research** – to improve our understanding of the drivers of health, well-being and disease, and to develop and evaluate a wide range of approaches to predict, prevent, diagnose and treat health conditions, both common and rare, across the entire course of people's lives.

Each one of these uses must flourish for our health and care system to be able both to provide the best care for individual patients and to understand and serve the health needs of the wider population. Each relies on access to different component parts of the same underpinning data, assembled and accessed in different ways. Further, although these uses are usefully described as distinct activities, it is important to recognise that there are substantial overlaps, or at the very least blurred margins, between the uses and between the people and organisations involved in them.

1.1 National health data for 67 million people in the UK

While our ability in the UK to use data optimally for each of the four uses outlined above is far from perfect, we have some precious assets that we must not lose sight of. Prominent among these is the potential to build a detailed, dynamic picture of the health status of the entire population of the UK, by linking different nationally collated sources of health data together. A crucial ingredient of this potential capability is our publicly funded NHS, established to provide healthcare for all 67 million of us across the four nations of the UK.[10] Over 98% of people in the UK are registered with the NHS through a general practice.[11] And, of the small percentage remaining who are not, many will have had contact with other parts of the health service (for example hospital emergency departments). Although under immense pressure and struggling to meet the demands of a rapidly ageing population in the post-Brexit, post-COVID era, the NHS continues to provide almost all healthcare for the entire population, remaining – for the most part – free at the

10 England 57 million, Scotland 5.5 million, Wales 3.1 million, Northern Ireland 1.9 million in 2021. See https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2021.

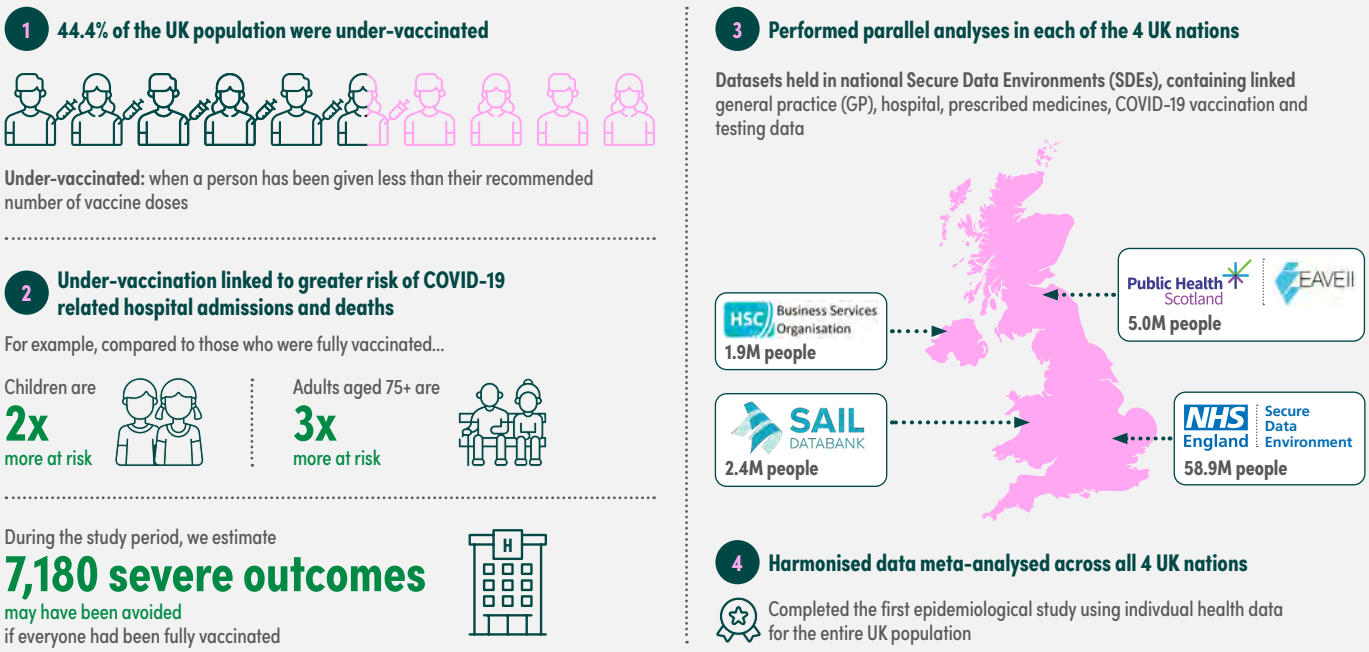
11 This 98% figure is cited frequently (e.g. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6929522/pdf/dyz034.pdf). It appears to be a reasonable estimate but it is difficult to find a clear source of evidence for it. Overall, general practice list sizes in fact exceed Office for National Statistics population estimates by a few percent. There are several potential explanations for over- and under-estimation of the number of patients registered with a general practice. For more detail, see: https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice/data-quality-statement; https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice/april-2021/spotlight-report-april-2021; https://commonslibrary.parliament.uk/population-estimates-gp-registers-why-the-difference/; https://publichealthscotland.scot/publications/general-practice-gp-workforce-and-practice-list-sizes/general-practice-gp-workforce-and-practice-list-sizes-2012-2022/; https://statswales.gov.wales/Catalogue/Health-and-Social-Care/General-Medical-Services/General-practice-population/patients-registered-at-a-gp-practice; https://www.opendatani.gov.uk/dataset/gp-practice-list-sizes.

point of delivery.[12] Researchers' ability to use UK-wide, whole-population health data was greatly enhanced during the recent COVID-19 pandemic. This was demonstrated recently in the first published analysis to include multiple linked sources of health data from almost everyone in England, Scotland, Wales and Northern Ireland in a UK-wide study of the rates, potential causes and consequences of being incompletely vaccinated against COVID-19 during 2022 (Figure 1.1).[13] Such insights are needed to inform delivery of the most effective and equitable healthcare for all of us. But, despite tangible progress, many more insights – and substantially greater benefit – can be generated from these already existing data. It is therefore essential that our national data capability is maintained and further improved beyond the COVID-19 pandemic. This should be used to address not only the consequences of COVID-19 but also the challenges of a health and care system in crisis, facing the ongoing 'pandemics' of dementia, heart disease, stroke, cancer, diabetes, arthritis, mental health, respiratory, eye and kidney diseases, and a plethora of other health conditions, both common and rare.

12 Office for National Statistics analyses found that total UK healthcare expenditure in 2023 accounted for 10.9% of gross domestic product. Of this, 82% was government-financed, 13.8% was out-of-pocket expenditure, 2.5% was voluntary health insurance, and 1.8% was non-profit institutions serving households and enterprise financing (https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthcaresystem/bulletins/ukhealthaccounts/2022and2023). A large proportion of privately funded healthcare spend is on elective procedures, with almost 25% of hip and knee replacements and almost 10% of cataract procedures funded privately in 2021/22 (see Nuffield Trust Explainer – *How much planned care in England is delivered and funded privately?*: https://www.nuffieldtrust.org.uk/resource/how-much-planned-care-in-england-is-delivered-and-funded-privately and updates from the Private Health Information Network: https://www.phin.org.uk/news/phin-private-market-update-december-2023).

13 See HDR UK COALESCE Consortium. *Undervaccination and severe COVID-19 outcomes*. Lancet 2024. https://www.thelancet.com/action/showPdf?pii=S0140-6736%2823%2902467-4.

Figure 1.1 Using multiple sources of linked health data from the whole UK population to understand COVID-19 vaccine uptake[14]



14 Adapted from an image created by the Usher Institute, University of Edinburgh.

Globally, the ability to use national whole-population health data is unusual. Where it exists (for example in several Scandinavian countries, Estonia and Israel), this is generally at a smaller scale (covering up to 10 million people, in comparison with the 57 million in England or 67 million UK-wide). Preserving and enhancing this uniquely large-scale, inclusive national capability across the UK is essential.

In each of the four nations of the UK, datasets derived from different parts of the healthcare system have, for decades, been collated at national (whole-country) level. These national population-wide data provide the essential building blocks for generating system-wide intelligence, based on data that includes – and so is relevant to – all people who use or have had contact with the NHS, whatever their age, sex, ethnic group, where they live, how rich or poor they are, and their previous and current mental and physical health. When linked (at person level) across different sources, accessed securely and analysed appropriately, these types of nationally collated data can be used for many beneficial purposes, for example:

1. To **describe, understand, plan and monitor healthcare provision** in different parts of the health service across each of the four nations, including their regions and other localities. This could include identifying and highlighting inequalities in healthcare provision and outcomes related to age, sex, ethnic group,

geographic location and a range of other socio-demographic factors, so that they can be properly understood and addressed.

2. To **identify and invite people for disease prevention programmes,** such as national screening (for example for various cancers) or vaccination programmes (for example for COVID-19, measles etc), to monitor uptake and to **follow the health** of those receiving them to assess their effectiveness in preventing disease in individuals and across the population.

3. To conduct **whole-population health studies** that link and analyse different sources of data, including – and so relevant to – everyone in the population. Such studies can improve our understanding of the causes of or risk factors for diseases, as well as ways to predict, prevent and treat health conditions that can affect people at different times during their lives.

4. To **identify and invite people to take part in population health or clinical research studies (including clinical trials)** occurring across one or more nations of the UK, and to follow the health of participants in such studies over time. These include studies investigating how our genes, lifestyle and environment cause different diseases, with the aim of developing new prevention and treatment strategies; testing new methods for earlier detection of diseases such as cancer or dementia; or assessing the balance of benefits and risks of new medicines for the prevention or treatment of diseases such as heart attacks, strokes or diabetes.

The ability to link national-scale health data underpins many of the UK's most prominent, widely cited and internationally leading successes in the life sciences, which showcase the UK's NHS, research community (both university and industry-based) and funders (both public and private) working together at their best. They have been essential for the successful delivery of large-scale clinical trials, such as the RECOVERY trial of treatments for severe COVID-19, influenza and community-acquired pneumonia[15] or the NHS-Galleri trial, evaluating a new blood test for the early detection of cancers.[16] They also provide essential data for following the health of research volunteers in UK population-based genomic medical research resources, such as UK Biobank,[17] Our Future Health,[18] and Genomics England,[19] which involve hundreds of thousands of participants and are used by tens of thousands of researchers to make discoveries to improve human health and well-being.

15 https://www.recoverytrial.net/.

16 https://www.nhs-galleri.org/.

17 https://www.ukbiobank.ac.uk/.

18 https://ourfuturehealth.org.uk/.

19 https://www.genomicsengland.co.uk/.

Chapter 2

Patient, public and health professional views on uses of health-relevant data

In this chapter

2.1 Views of patients and the public across society

Most people support the use of their health data to benefit society. Over the last 10–15 years, surveys, qualitative studies, workshops, consultation discussions, citizens' juries and other exercises have gathered and reported the views of patients and the public on access to and uses of health-relevant data. These have been conducted or commissioned by various national public bodies, academic research groups, and independent social science and policy organisations.

The organisation Understanding Patient Data[20] has produced a very helpful summary of the most relevant work, covering the 10-year period to 2021.[21] This focuses on views from across society about the uses of routinely collected data from health and care systems, where explicit consent is not sought from each person for the use of their data. Additional relevant work since then includes: a survey in 2022 led by NHS Digital aiming to better understand public views on uses of NHS data after the planned *General Practice Data for Planning and Research* data collection was paused in late 2021;[22] a survey and deliberation exercise in 2022 commissioned by NHS England to help understand patient and public views of its principles for access to NHS data by commercial organisations;[23] and a survey and series of workshops commissioned by Understanding Patient Data in 2024 exploring attitudes about, awareness of and support for different uses of NHS data.[24]

While there is no such thing as an average view, several consistent themes emerge from this large body of work, summarised in Table 2.1. Similar themes arose in our discussions with organisations representing patients and the public and in our public workshops (see Appendix 5 for a summary of our workshop findings).

20 See https://understandingpatientdata.org.uk/.

21 See https://understandingpatientdata.org.uk/how-do-people-feel-about-use-data.

22 See https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/general-practice-data-for-planning-and-research/gpdpr-programme-reports-and-publications/public-survey-summary-report.

23 See https://transform.england.nhs.uk/key-tools-and-info/centre-improving-data-collaboration/guide-to-effective-nhs-data-partnerships/#4-how-to-obtain-fair-value-for-the-public-from-data-partnerships.

24 See https://understandingpatientdata.org.uk/public-attitudes-patient-data-planning-and-population-health.

Table 2.1 Summary of patient and public views from the last 15 years on uses of data from health and care systems where explicit consent is not sought from every person for use of their data

Data sharing for direct care

- The great majority of people (well over 90%) welcomes and expects sharing of health and care records for their direct care.[25]

Public benefit for data uses beyond direct care[26]

- Most people (80–90% or more in most surveys) readily accept – and many expect – the use of patient-level data for reasons beyond their direct care, for example to plan services and to better understand and treat diseases, provided this is clearly for public benefit.
- People do not feel that uses of data need to remain close to the original purpose of collecting the data to bring public benefit.
- People want benefits arising from data use to be distributed widely and equitably across all groups in society. They see value in uses that could benefit small and large numbers of people (for example those with rare as well as those with common health conditions).

Transparency and accountability

- People want clear information on who has access to their data, for what purpose and why. Such transparency is essential to avoid suspicion and mistrust.
- People want decisions about access to data to go through a transparent process with external oversight and clear accountability.

25 This reflects my own experience as a doctor as well as that of my many health professional colleagues. Patients are frequently surprised or disappointed when relevant health information about them from different parts of the health system is not readily accessible to health professionals involved in their care. "Do you not know that from my GP records?" is frequently asked by patients in hospital emergency department, outpatient clinic and ward settings. Occasionally, patients request restrictions on the sharing of their data (e.g., if worried about wider sharing of asylum or immigration status or disclosure of a sensitive diagnosis such as a sexually transmitted disease). But the vast majority expect their personal health information to be available (with appropriate security and confidentiality) across the system and are rightly concerned for the quality and safety of their healthcare if this is not so.

26 The National Data Guardian has produced guidance on evaluating public benefit for uses of health and care data beyond individual care. See: https://www.gov.uk/government/publications/what-do-we-mean-by-public-benefit-evaluating-public-benefit-when-health-and-adult-social-care-data-is-used-for-purposes-beyond-individual-care.

Public involvement and engagement

- People feel that the public should have a say in how health and care data are used.
- Generating and maintaining public confidence and trust needs engagement with and involvement of people from a cross-section of society in ongoing, open discussions about how data are collected, stored, accessed and used.

Trust and confidence in different uses and users

- Public support for use of health and care data depends on the organisations holding, managing, accessing and/or using these data being competent and perceived as competent (particularly as regards data security and privacy) and having the right motivations (i.e. to achieve public benefit).
- People have high levels of trust in healthcare professionals (especially GPs) and public sector researchers accessing and using patient data for approved purposes. But there are concerns about access and use by commercial organisations and there is resistance to patient data being used for insurance or marketing purposes.
- When provided with more information about the role of commercial companies in improving healthcare, for example developing new ways to prevent, diagnose and treat disease, people are far more supportive, preferring that commercial companies access data than society miss out on the research benefits.
- Levels of trust and confidence in use of health and care data are lower among ethnic minority and more deprived groups.
- People say that they would have higher trust in uses of their data if they could more readily access their own records and correct inaccuracies.

Keeping data safe

- Data security and privacy are important: people feel that their personal health data should be treated with the utmost care but recognise the potential to bring public benefit if this happens.
- People also want safeguards to protect society from misuses of data that could cause harm.

Commercial use and benefit share

- People feel that profitable uses of data should have particular scrutiny to ensure public benefit.
- They also feel that the NHS should benefit if commercial organisations produce something of value, for example by giving the NHS a preferential rate for any product or service developed with NHS data, or unlimited access to new knowledge and insights arising from a company's work with NHS patient data.
- People have greater trust in the potential for public benefit and in the safe use of their data if commercial organisations are involved through a partnership with the NHS.

Public understanding

- The public has limited awareness and understanding of:
  - the potential uses of health and care data beyond individual care;
  - how commercial companies can bring public benefit from access to patient data;
  - current data collection processes, data safeguards, opt-out processes, and the processes of data de-identification, de-personalisation or complete anonymisation.
- This is partly because the words used to describe patient data and their uses can be complex and confusing; understanding improves when clear, non-technical language is used.
- However, it is also important to recognise that not everyone wants to know or understand more about these issues.

Many other sources of administrative data from across government departments and public services (such as social services and education) are relevant to health. So, an understanding of public attitudes to data sharing more broadly is relevant. The Office for National Statistics (ONS) has provided a very helpful summary of what is known about this from a wide range of public engagement exercises conducted over the last 15–20 years.[27] The key issues and public attitudes are very like those relating to data from the healthcare system and have not changed substantially over the last 15 or more years. They indicate that public trust for uses of administrative data from different sources depends on:

- demonstrable **public benefit** from data uses;
- **transparency** about how, by whom and for what purpose data are used;
- clarity on robust mechanisms to **keep data secure** and to **maintain people's confidentiality and privacy**; and
- **independent oversight** of data uses.

The ONS itself is highly trusted: when asked, 85–90% of people report trust in the ONS and the independent statistics it produces.[28] This legitimises the crucial national role that the ONS plays in enabling the collection, linkage and analysis of – and safe researcher access to – multiple sources of data relevant to health (see section 3.2.3).

2.2 Views of participants in specific research studies or resources

It is important to recognise the distinction between the views of patients and the public across wider society and the views of people who have explicitly agreed to take part in a specific research study (such as a population-based cohort study or a clinical trial of a new treatment) or research resource (see section 3.3.1 and 3.3.2). The second group comprises people who have actively chosen to be research participants, often making a considerable commitment (for example by completing questionnaires, undertaking various measurements and tests, providing samples for analysis, or taking a study drug or placebo in a randomised clinical trial). They will generally have explicitly consented to take part. In addition, participant information, consent and study processes will have been carefully scrutinised and approved by an independent research ethics committee.

Most research studies and resources of this type actively engage with their participants during their follow-up (through providing information at study visits, issuing newsletters, sending invitations to participate in additional data collection exercises, and so on), which may continue for years or even decades. These research participants have usually volunteered to take part on the understanding that their data and/or samples will be used to address various research questions. They have often provided specific consent for data from their NHS and other health-relevant records to be integrated into the research study or resource to enhance its research value. If they learn that this has not happened or is not happening, they may quite rightly be surprised, annoyed or even outraged. Unfortunately, data integration does

27 https://www.ons.gov.uk/aboutus/usingpublicdatatoproducestatistics/peoplesattitudestodata/whatweknowfromengagingwiththepublicondatajune2023.

28 See https://natcen.ac.uk/sites/default/files/2022-12/NatCen_Public-Confidence-in-official-statistics_2021.pdf.

not always happen, particularly in the case of linkages to general practice data. Some of the reasons are outlined in the following section.

2.3 Views of healthcare professionals

Many studies provide information about the views of patients and the public on uses of health data (section 2.1), but the views of healthcare professionals have been less widely studied. While some GPs are at the forefront of efforts to ensure the use of data for patient and public benefit, others are more circumspect. Understanding the views of GPs is crucial, given that:

- they are currently the legal data controllers of the health records generated within each general practice;
- access to and linkage of primary care data was the highest priority and unmet need for almost all stakeholders we consulted with;
- access to and linkage of the structured, coded component of general practice electronic health records is technically and administratively straightforward (see section 3.1.2), so barriers to sharing are likely to lie elsewhere.

Two recent pieces of work, conducted on behalf of Understanding Patient Data and NHS Digital, explored the views of general practice staff, mainly GPs but also practice managers, nurses and allied health professionals.[29] Both were conducted shortly after the *General Practice Data for Planning and Research* programme stalled in 2021. Both conducted surveys that between them obtained complete responses from less than 350 general practice staff. This is a tiny fraction of the many tens of thousands of GPs and other staff who work in general practices across the UK. Both found the respondents to be broadly supportive of the sharing of healthcare records for the clinical care of individual patients as well as for planning and research. However, support for and confidence in data sharing was generally lower among general practice staff than among patients and members of the public. The confidence of practice staff was higher for sharing data within their primary care network than with less familiar NHS and non-NHS organisations beyond the practice and primary care network. For example, of 111 general practice staff (92% of whom were general practitioners) responding to the survey conducted for Understanding Patient Data, 79–81% reported being comfortable with the sharing of data for care, planning or research purposes across their primary care network. Percentage figures fell to 51–65% for sharing more widely across the NHS, and to 18–28% for sharing more widely beyond the NHS. Neither survey asked about uses of data held within secure data environments. Interestingly, both studies found greater support for data sharing for planning and research than for clinical care.

29 See NHS Digital *GP Staff Survey Summary Report*, 2022: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/general-practice-data-for-planning-and-research/gpdpr-programme-reports-and-publications/gp-staff-survey-summary-report-general-practice-data-for-planning-and-research-gpdpr; and work commissioned from Mott MacDonald by Understanding Patient Data, *Primary Care Professionals' Attitudes to Data Use*, 2022: https://understandingpatientdata.org.uk/sites/default/files/2022-03/Primary%20care%20professionals%27%20attitudes%20to%20data%20use%20.pdf.

Several factors influence these views:

1. Practice staff take their contractual, legal and professional responsibilities very seriously. Concerns about the time needed to engage with data sharing and the potential legal liability (for example for inadvertently breaching data protection laws or obligations under the common law duty of confidentiality) are a barrier to data sharing. Mechanisms to ensure that streamlined, trustworthy, compliant systems for data sharing are available to GPs to minimise the time burden and provide greater assurance would help. Some staff favour a centralised data collection process whereby a central NHS team or organisation has responsibility for determining when and how patient data will be shared for research and analysis purposes.[30] Some feel that better quality information is needed on how data are used and kept secure, and about the benefits of data use.
2. For many practice staff, **use of data for analysis or research is not seen as a core activity**, especially when the daily business of managing the immediate needs of patients is all-consuming in the face of **very limited time and resources.**
3. The **anxieties** of practice staff **about potential uses of data for performance management** may be a further barrier to some data sharing.
4. **Some practices have a culture of data sharing**, with good understanding of how data is used safely and the associated benefits. For example, around 30% of practices provide data to the Clinical Practice Research Datalink.[31] This allows them to contribute to research; earn extra income from simple questionnaires and clinical studies; receive quality improvement reports; and accrue evidence for professional appraisal and revalidation.
5. **Financial incentives** may be an important motivator for some practices (for example these were successful for the Quality Outcomes Framework) and may encourage data sharing for research in some cases.[32]

In summary, identifying positive benefits and providing secure frameworks and processes for data sharing that minimise actual and perceived risks appear to increase the willingness of primary care health professionals to advocate for and support the sharing of, access to and use of health data.

2.4 Perceptions and misperceptions of the risks and benefits of data uses

Understanding attitudes to the risks versus benefits of sharing and access to health data is important. A detailed discussion of understanding, perceptions and misperceptions of the risks (for example of data misuse or of a data breach[33]) is beyond the scope of this review. However, there are some specific issues to highlight.

30 Views on reducing the decision-making responsibilities of general practices do vary, however. E.g., our discussions highlighted the fact that some GPs may be reluctant to relinquish decision-making responsibility and power to a central NHS body without demonstrable understanding and support from the UK government and the DHSC for the challenges GPs face in providing quality patient care with limited resources.

31 See https://www.cprd.com/.

32 We note, however, that financial incentives alone are not sufficient to enable data sharing by all practices. E.g., the experience of UK Biobank in seeking permission from GPs to obtain data for its 500,000 participants (who have provided explicit consent for this) has been that financial incentives alone do not overcome other concerns, such as data controllership legal liability. See https://www.ukbiobank.ac.uk/using-gp-data-of-uk-biobank-participants and https://www.ukbiobank.ac.uk/media/qjjpzp4e/october-ukb-obstacles-to-obtaining-coded-gp-data-2710-003.pdf.

33 A personal data breach means a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data (e.g. if a laptop containing personal data is lost or stolen, a letter is sent to the wrong address, someone without proper authorisation accesses data or passes it on, or data are hacked). A breach may be accidental or deliberate.

First, concerns about lack of transparency, potential for data misuse and risk of data breaches are sometimes conflated. This is unhelpful, since understanding concerns and seeking solutions to them is only possible if each concern is addressed separately. Examples are the objections to the stalled General Practice Data for Planning and Research Programme,[34] each of which needs to be addressed, including those related to:

- the need for clearer information for patients, the public and professionals (i.e. transparency);
- concerns about the way the data might be used (for example for profit or performance management); and
- concerns about the potential for inadvertent or deliberate re-identification of people in the data.

Second, news stories about data breaches relating to health data can create disproportionate concern. It is important to recognise that the majority of known health data breaches have related to the inadvertent loss, sharing of, or deliberate illegal access to **directly identifiable data** in healthcare systems.[35] These occurrences are of course very concerning and need to be taken very seriously, with ongoing monitoring, data security reviews, risk mitigation strategies, and penalties where appropriate. The risk of such data breaches is small: in the last five years, on a background of over 500 million patient interactions each year in the NHS, fewer than 2000 health data breach incidents per year were reported to the information commissioner. However, it is very difficult to quantify the impact of these incidents on the people whose data were involved.

The risk of deliberate re-identification when **de-identified data** are accessed by approved researchers (especially if this is within a secure research environment) is extremely small. There are criminal penalties for deliberately trying to re-identify someone from such data without permission. The implementation of several layers of safety through the Five Safes Framework (see section 5.5) means that data breaches of this type in such settings are so rare that this risk tends to be illustrated with theoretical examples,[36] or through reference to incidents involving directly identifiable rather than de-identified data. We were not able to find any examples of deliberate re-identification of individuals by researchers, despite thousands of researchers accessing millions of de-identified records over at least two decades, for example via approved access to the SAIL Databank,[37] the Scottish National Data Safe Haven,[38] the NHS England Secure Data Environment,[39] the Office for National Statistics Secure Research Service,[40]

34 See https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/general-practice-data-for-planning-and-research/about-the-gpdpr-programme.

35 See https://understandingpatientdata.org.uk/weighing-up-risks#data-breaches-in-the-health-sector.

36 The potential to interrogate the private health records of former prime minister Tony Blair is often used as an example. In a debate on patient data in the House of Commons in June 2021, David Davis said: *"Take Tony Blair, who was widely known to have developed a heart condition, supraventricular tachycardia, in October 2003. He was first admitted to Stoke Mandeville and then rushed to Hammersmith. One year later, in September 2004, he visited Hammersmith again for a corrective operation. Even the name of the cardiologist is in the public record. A competent researcher would make very short work of finding such individual records in a mass database."* (https://hansard.parliament.uk/commons/2021-06-24/debates/2FA13B90-5377-4E73-A941-80F6A536B560/UseOfPatientData). In theory, a researcher accessing de-identified health records from the population of England could search among millions of sets of patient records to find those of Tony Blair and discover some additional information about him that is not already in the public domain. However, in practice, no data breaches of this sort have been reported.

37 See https://saildatabank.com/.

38 See https://www.nhsresearchscotland.org.uk/research-in-scotland/data/safe-havens.

39 See https://digital.nhs.uk/services/secure-data-environment-service.

40 See https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice.

The Northern Ireland Honest Broker Service,[41] UK Biobank,[42] or the Clinical Practice Research Datalink.[43]

Finally, when it comes to health data access for patient and public benefit, we tend as a society (certainly in the UK, but also in many other countries) to focus far more on the risks of – and liability for – data misuse or data breaches than on the risks of not providing access to data. Understanding Patient Data makes every effort to provide balanced information,[44] considering not only the risks of sharing but also the consequences of not sharing health data, commenting that:

"The failure to record, link and share data can negatively impact patient care, and waste scarce resources. For example, looking at patterns in data is essential to monitor the long-term safety of drugs and treatments, and to identify adverse side effects as quickly as possible. Without effective use of data, services are not improved and patients will suffer."

Keeping these consequences in mind may help to drive appropriate sharing and uses of data for public benefit. This could be encouraged by setting clear expectations, providing incentives, or offering reassurance and practical protection against liability (for example for GPs). Instead of asking the question:

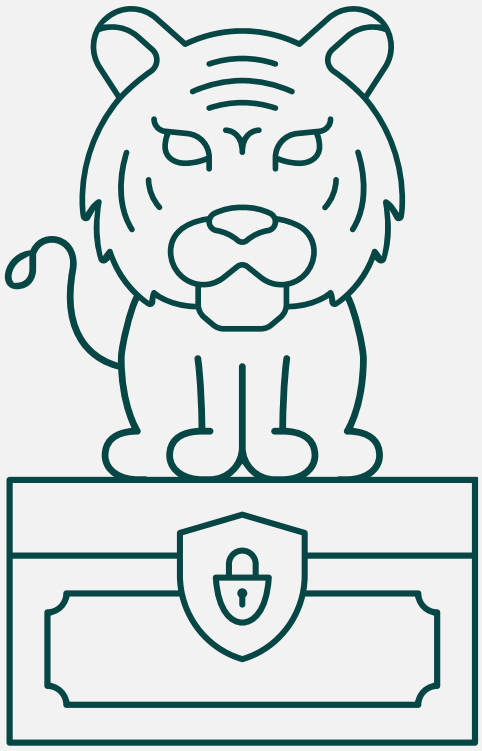
"What is the best way of ensuring that the data we are responsible for are used as widely as possible to maximise the benefit for patients and the public, while preserving privacy and maintaining security?"

data custodian organisations all too frequently focus instead on asking:

"How can we ensure that we protect the data that we are responsible for to minimise privacy and security risks?"

The answer to the second question is of course easy, as shown in Figure 2.1, but ultimately unhelpful as it prevents the realisation of patient and public benefit.

Figure 2.1 Data overprotection?



The logical answer to the question ***"How can we ensure that we protect the data that we are responsible for to minimise privacy and security risks and our liability?"*** is to lock the data away in a well-guarded location that no-one can access.

41 See https://bso.hscni.net/directorates/digital-operations/honest-broker-service/.

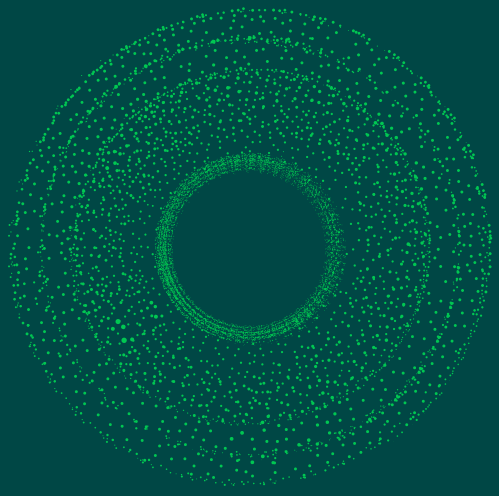
42 See https://www.ukbiobank.ac.uk/.

43 See https://www.cprd.com/.

44 https://understandingpatientdata.org.uk/weighing-up-risks.

Chapter 3

Sources of health-relevant data across the UK



In this chapter

The UK has abundant sources of health data that could be used by analysts, researchers and policymakers to improve people's health and benefit society. Here we review how, where and why data relevant to health are generated. We also touch on how these data are – or could be – used for a range of purposes that aim, ultimately, to benefit individual patients, their families and carers, and/or to improve the health and wellbeing of the wider public.

We can think of health-relevant data in several broad categories:

1. **Data from the healthcare system**, mainly arising from day-to-day activity in the NHS;
2. **Data relevant to health from other administrative activities and settings**, often related to local, regional or national government, and arising from day-to-day activity beyond or outside the health and care system;
3. **Data collected specifically for health research studies**, usually involving the recruitment of people who explicitly agree (consent) to take part in a particular study;
4. **Data relevant to health generated through environmental monitoring**, for example of the weather, climate, or air pollution;
5. **Data relevant to health generated by people as part of their day-to-day lives**, for example via mobile phone apps or wearable monitoring devices (for example an Apple watch or Fitbit).

As this review focuses on data relevant to health, we provide greater breadth of coverage, depth and detail on the first category (data from the healthcare system). The remaining categories are crucial in completing the overall picture of health-relevant data across the UK. We do not attempt to map or describe these comprehensively; rather, we explore some prominent examples and use these in later discussions of the barriers to – and recommended solutions for – the uses of health-relevant data for patient and public benefit.

In Chapter 1, we emphasised the critical importance of national-scale data resources and their huge potential to help tackle not only national issues but also many regional and local ones. If managed and used well, such national-scale resources have huge potential to reduce duplication of effort in the processes of collecting, curating, storing, accessing and analysing data, bringing economies and efficiencies of scale. Because of this, we particularly highlight sources of data that have national geographic coverage, that have already been collated at national level, or that have the greatest potential to scale nationally. These tend to be types of data that are highly structured, often coded, and in general not especially high volume (although there are some notable exceptions to this). Box 3.1 provides a brief explanation of the concepts of structured, unstructured and coded data, while Box 3.2 provides some background information on data volumes.

Box 3.1 Explaining structured, coded and unstructured data

Structured data

These are data that can be represented in the form of a spreadsheet (or similar): column headings describe each data item, and each row represents a specific person or the record of a health event relating to a person. Structured health data might include data items such as the name, date of birth and NHS number of the people included in the data; their characteristics, such as sex or ethnicity; the dates of various health events (for example date of diagnosis of a health condition or medical procedure) and the diagnoses or medical procedures that occurred on these dates.

Coded data

Structured data are often coded. This means that a number or a code is used to describe each item of data. For example, female might be coded as F and male as M. Numbers are often used to code data items (with a key explaining what each number means for that item). For example, the broad ethnic group categories of White, Black, Asian, Mixed and Other could be coded as 1,2,3,4, and 5 respectively. In health records, nationally and internationally recognised coding systems are used to describe, in a consistent way, health conditions, medications, operations, symptoms, tests and other information recorded by healthcare staff. Such coding makes it easier to analyse the records of large numbers of people together, for example to look at the trend in occurrence of a specific health condition over time, or to assess the relationship between ethnic group and a particular health condition.

Unstructured data

These are data such as imaging scans, notes made during hospital ward rounds, or letters written from one doctor to another describing their assessment of a patient. These days, unstructured data are usually recorded in electronic form but they cannot be represented in spreadsheet format. Unstructured data tend to occupy more space in computer systems (that is, they have greater volume) than structured data. It is possible to derive structured data from unstructured data, to make analysis of the data easier. For example, information can be extracted from medical notes on the diagnoses or procedures and the dates of these. This information can then be coded (for example by using national and international diagnosis and procedure coding systems) and included in structured datasets such as hospital episode statistics. This coding may be done manually (for example by hospital coding clerks), but increasingly it is possible to use automated methods to increase the speed and consistency of these types of tasks.

Box 3.2 Explaining data volumes

Data files occupy space in computer systems. The space occupied is referred to as data volume or data storage volume. The volume of data in a single file or file system can be described in terms of units called bytes. Data volumes can be very large, particularly when it comes to complex, unstructured data:

- Kilo means 1,000; a kilobyte is one thousand bytes
- Mega means 1,000,000; a megabyte is one thousand kilobytes
- Giga means 1,000,000,000; a gigabyte is one thousand megabytes
- Tera means 1,000,000,000,000; a terabyte is one thousand giga bytes
- Peta means 1,000,000,000,000,000; a petabyte is 1,000 terabytes
- Exa means 1,000,000,000,000,000,000; an exabyte is 1,000 petabytes
- Zetta means 1,000,000,000,000,000,000,000; a zettabyte is 1,000 exabytes

Some examples of the data volumes of familiar things give a sense of what these quantities mean:

Item	Approx. data volume
1 letter of text	1 byte
1 paragraph of text	1 kilobyte
1 book (with around 200 pages)	1 megabyte
1000–2000 books	1 gigabyte
250,000 (quarter of a million) songs in MP3 form	1 terabyte
745 million floppy disks	1 petabyte
12 billion DVDs or 16 trillion MP3 songs	1 exabyte
All data generated worldwide in 2016	1 zettabyte
All data generated worldwide in 2021	79 zettabytes

Some examples of the data volumes of health datasets, data collections or resources are shown below:

Item	Approx. data volume
One person's full genetic sequence[45]	30 gigabytes
All structured, coded general practice data for all 57 million people in England	3.6 terabytes
All NHS Scotland radiology imaging data from 2008–2018	3 petabytes
All data held by UK Biobank in 2024 for 500,000 participants	30 petabytes
All data in the English National Genomic Research Library by end 2022	65 petabytes

45 This refers to short-read sequencing, which generates less data of lower volume than long-read sequencing.

3.1 Data from the healthcare system

3.1.1 A complex and evolving system

Every day, healthcare staff record data (i.e. information of various types) about the patients they care for. While some of this information is still recorded on paper, it is increasingly entered into and stored electronically in computer systems. The primary purpose is to record and preserve important administrative and clinical details about each encounter with or about a particular patient, for example when a doctor or nurse discusses with a patient their diagnosis, tests, treatments or possible future health outcomes. This record is so that each patient's ongoing and future care is informed by relevant information about their past and current health (dates of health service attendances, appointments and admissions, symptoms, signs, diagnoses, treatments, operative procedures and so on).

To provide the best care, health and care professionals need to access existing information and to record new information about each patient. This information needs to be up to date and available when healthcare discussions, decisions or activities are taking place. The information may be held in multiple different computer systems used by different parts of the healthcare system. This situation arises because, during their lives, many people will receive care in one or more hospitals (for example in a maternity unit when they are born, in an accident and emergency department for an injury, or in a specialist clinic or hospital ward for assessment and treatment of a health condition) as well as from their general practice and other community health services (for example pharmacy, optician, dentist).

As we move rapidly towards a paperless NHS, the healthcare system is increasingly adopting computer systems that are more sophisticated. These aim to provide secure access to the right information at the right time for healthcare professionals and the patients they care for. However, although most patients might hope and expect that the doctor, nurse or other healthcare professional caring for them can quickly access all the information about them that they might need, this is often not the case.[46] Similarly, one might expect that it should be reasonably straightforward for the NHS to gather the relevant data that it holds about the health and healthcare of groups of patients or of larger regional or national populations to support healthcare planning or health research. But several inter-related layers of complexity can make this very challenging in practice. These include:

1. **Organisational complexity**. While often thought of as a single healthcare system, the NHS consists of many national, regional and local organisations. These have various organisational labels, definitions, groupings, roles and responsibilities that change over time, sometimes alarming rapidly. This can be confusing for patients, the wider public, NHS staff and others.[47,48] The many different types and sources of health data are not the responsibility of a single NHS organisation; rather, data custodianship is distributed across many organisations. This fragmentation can result in unhelpful competition and lack

46 See https://rorycellanjones.substack.com/p/after-the-fall-the-investigation for a salutary tale about how flawed and poorly interoperable NHS computer systems contributed to inadequate NHS care of former BBC technology correspondent Rory Cellan-Jones, following a fall in late 2023.

47 See https://www.nuffieldtrust.org.uk/features/nhs-reform-timeline for a brief historical perspective of NHS reform from the 1940s to 2022.

48 Writing as MD in *Private Eye* in July 2023, Dr Phil Hammond (NHS GP, broadcaster, comedian and commentator on health issues in the UK), described the last 30 or so years of reforms in the NHS in England as follows: *"In my professional lifetime, the NHS had … the purchaser-provider split, GP fundholding, competitive tendering, Trusts, Foundation Trusts, Primary Care Trusts, Health Improvement Plans, the National Institute for Health and Care Excellence, the Healthcare Commission, the Commission for Health Improvement, Practice-Based Commissioning, Polyclinics, NHS Commissioning Boards, NHS England, Monitor, Healthwatch, the Care Quality Commission, GP Pathfinder Consortia, Clinical Commissioning Groups, Clinical Support Units, the NHS Trust Development Authority, Public Health England, NHS Improvement, Sustainability and Transformation Plans, Primary Care Networks, Integrated Care Systems and Integrated Care Boards."* See https://www.reddit.com/r/nhs/comments/14zuj8t/private_eye_md_on_the_nhs/ for the full article.

of trust between different parts of the NHS, compromising effective data access and sharing across and between organisations. In addition, devolved responsibility for health and care means that there are differences in health and social care policies and their organisational structures between the UK's four nations.[49] These include differences in national and regional data collections, which contribute to the challenges of conducting consistent analyses of health data across all four nations or of making unbiased comparisons between countries.

2. **Computer system complexity**. Many different computer systems are used by the many NHS organisations. Some are developed by NHS organisations themselves, but most are provided through contracts with multiple commercial computer and software system suppliers.[50] The introduction, development, and integration of these different systems over the years has varied – and continues to vary – both geographically (between and within the four nations of the UK) and across different parts of the health service (for example general practices, hospitals, diagnostic laboratories, radiology departments conducting X-rays and scans, and so on). The result of this complexity is that the various computer systems are not fully interoperable. In other words, they do not readily support sharing of or access to information with other systems.[51] This often makes it difficult for health and care professionals to access all the information that could and should be used for individual patient care. It can also make it difficult to pull together information about multiple patients or populations to plan or deliver services or to support medical research.[52]

3. **Transactional complexity**. Achieving interoperability between computer systems is not just about resolving technical computer system interoperability challenges. It also needs robust, transparent contractual agreements and trusted partnerships between the many NHS and commercial organisations involved. Experience and common sense suggest that the effort and cost (much of which is borne by the taxpayer) of managing transactions across multiple organisations (especially competing ones) increases as the number of these organisations increases.

4. **Legal and regulatory complexity**. The combination of legislation and common law relevant to the use and sharing of confidential health data for different purposes is complex. This complexity is exacerbated by differences in the common law duty of confidentiality and processes for enabling lawful use of confidential patient information between the four nations of the UK, as well as differences in the interpretation and application of data protection legislation and common law requirements by the many different NHS and non-NHS organisations that collect, hold, use and provide health data. A consequence of this complexity and variability is a tendency towards decision-making on data sharing and access that emphasises the avoidance of risk over the realisation of benefits for patients and the public. The more organisations involved in these decisions, the greater this tendency.

49 For further detail, see: https://www.instituteforgovernment.org.uk/explainer/devolution-and-nhs#footnoteref32_08nt6g6; https://www.instituteforgovernment.org.uk/report/devolved-public-services; https://www.nuffieldtrust.org.uk/research/integrating-health-and-social-care-a-comparison-of-policy-and-progress-across-the-four-countries-of-the-uk#report-overview.

50 E.g. see https://digital.nhs.uk/services/digital-services-for-integrated-care; https://www.england.nhs.uk/hssf/supplier-lists/.

51 Interoperability between different computer systems requires the development and maintenance of so-called Application Programme Interfaces or APIs. The greater the number of computer systems, the greater the number and complexity of API solutions required to enable interoperability. For information on NHS England's current API platform, policy and roadmap see https://digital.nhs.uk/services/api-platform.

52 https://www.england.nhs.uk/digitaltechnology/digitising-connecting-and-transforming-health-and-care/.

3.1.2 General practice data

Broad health information about almost everyone in the UK

Around 98% of people living in the UK are registered with one of over 8000 NHS general practices across the UK (around 6500 in England, 1000 in Scotland, 400 in Wales and 300 in Northern Ireland).[53] GPs, practice nurses and other practice staff enter information about their patients into general practice computer systems. This includes information recorded during clinical consultations about patients' symptoms, signs, observations, measurements, suspected or known diagnoses, medicine prescriptions, treatments received and specialist referrals (for example to hospital specialists). It also includes information derived from correspondence received (electronically or by postal mail) from healthcare professionals in other healthcare settings (for example hospital-based specialists). Much of the information in general practice computer systems is captured and stored as structured, coded data, using clinical coding systems (see Box 3.1).[54] Additional, often more detailed information is captured and stored as unstructured free text, for example notes made during consultations or correspondence to and from specialist services.

For many people, their general practice electronic health record contains the most comprehensive health information about them held in any single healthcare computer system. This is because the record includes information from their general practice care as well as information incorporated into their general practice record from other parts of the healthcare system. However, some patients receive a large proportion of their care and monitoring in specialist hospital settings (for example those with severe, prolonged and/or rare conditions). For them, the most detailed and up-to-date information about their health will often be held in the hospital computer system or systems, with summary information incorporated intermittently into their general practice electronic health record.

53 This 98% figure is cited frequently (e.g. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6929522/pdf/dyz034.pdf). It appears to be a reasonable estimate but it is difficult to find a clear source of evidence for it. Overall, general practice list sizes in fact exceed ONS population estimates (e.g. by up to 6% in 2019). There are several potential explanations for over- and under-estimation of the number of patients registered with a general practice, but no information that we could find in the public domain that quantifies or resolves these. For more detail, see: https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice/data-quality-statement; https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice/april-2021/spotlight-report-april-2021; https://commonslibrary.parliament.uk/population-estimates-gp-registers-why-the-difference/; https://publichealthscotland.scot/publications/general-practice-gp-workforce-and-practice-list-sizes/general-practice-gp-workforce-and-practice-list-sizes-2012-2022/; https://statswales.gov.wales/Catalogue/Health-and-Social-Care/General-Medical-Services/General-practice-population/patients-registered-at-a-gp-practice; https://www.opendatani.gov.uk/dataset/gp-practice-list-sizes.

54 Such as SNOMED-CT or Read for clinical terms and Dictionary of Medicines and Devices (d m + d) codes for medication and device prescriptions (https://digital.nhs.uk/services/terminology-server#content-included-in-the-terminology-server).

Long established electronic patient records in just a few computer systems

General practices have used computer systems to record patient information electronically since the 1980s. Currently, just three commercial companies provide the main general practice computer systems for the vast majority of the UK's 8000 general practices, and in each UK nation two out of these three cover almost all practices: EMIS[55] or TPP[56] cover almost all general practices in England; and EMIS or Cegedim[57] cover practices in Scotland, Wales and Northern Ireland.[58] Over many years, these computer system suppliers have developed technical solutions (or used solutions developed by others) to share data securely with other NHS and non-NHS organisations, including national NHS bodies. These include solutions to: (i) transfer a single patient's general practice record when they move from one practice to another one;[59] (ii) transfer general practice records for groups of patients from one or more general practices to another organisation; (iii) provide access to data within the data centres of the general practice computer system suppliers.

Sharing and linking general practice data at national scale is technically straightforward

The limited number of general practice computer systems, together with these established data sharing mechanisms, mean that enabling access to general practice data covering all or defined subsets of people registered with a GP in each UK country is technically and organisationally reasonably straightforward. It is worth noting that if the number of commercial general practice system suppliers were to increase in the future (which is possible),[60] this would add complexity, making this population-wide capability more difficult to achieve. This capability applies particularly to the structured coded component of general practice records (Box 3.1). By modern standards, this does not represent a particularly large volume of data (Box 3.2), even when considering the records of tens of millions of people, for example the whole population of England.[61] There are, however, other challenges of access to general practice data, and of their linkage to other data sources. These have been touched on in section 2.3 and are explored further in sections 6.3.2 and 7.2.1.

55 https://www.emishealth.com/: Egton Medical Information Systems (EMIS), providing the EMIS Web and EMIS PCS (primary care system) computer systems, used by almost all non-TPP practices in England and a significant proportion of practices across the devolved nations of the UK.

56 https://tpp-uk.com/: The Phoenix Partnership (TPP), providing the SystmOne general practice computer system, used by >2,700 practices in England.

57 https://www.cegedim-healthcare.co.uk/ providing the Vision 3 general practice computer system, previously used by a large proportion of English practices and currently used by non-EMIS practices in the devolved nations of the UK.

58 These companies provide their services under the contractual frameworks of the relevant national NHS organisations in each of the UK's four nations.

59 In England, this often uses a secure electronic transfer system called GP2GP, although this only works for transfers of data between practices within England, not for transfers between England and any other UK nation (where paper record print-outs are required): https://digital.nhs.uk/services/gp2gp#top; https://pcse.england.nhs.uk/help/medical-records/records-movement-pcse-online; https://www.nhsinform.scot/care-support-and-rights/nhs-services/doctors/transfer-of-your-gp-health-records/#moving-your-health-records.

60 E.g., see https://digital.nhs.uk//services/digital-services-for-integrated-care/gp-it-futures-systems for information on current and future plans for general practice IT systems and services.

61 By way of illustration, the estimated data volume of the structured, coded general practice data for the entire English population (almost 60 million people): (i) represents only a small proportion of the record level health data from multiple sources already collected and held by NHS England); (ii) comprises only around 0.01% of the total volume of data held on 0.5 million people in the UK Biobank research database (which holds over 30 petabytes of data).

3.1.3 Data from community-based health services other than general practices

Other community-based health services (apart from general practices) collect information about healthcare interactions between healthcare professionals and patients, recording data using a range of different commercial computer and software systems. These include dentists, high street opticians and a number of other community health service providers. Data from community pharmacies and from the community-based services of mental healthcare and maternity care service providers are covered in specific sections: Data on prescribed and dispensed medicines (section 3.1.5), Mental health data (section 3.1.9) and Maternity data (section 3.1.10).

Limited national-scale data from NHS-funded community dental and eye care

Most of the detailed health data recorded by dentists and high street opticians (including, for example, retinal images from retinal photography and dental X-rays) are kept within their commercial computer systems and are not accessed or analysed outside of these systems for wider benefit (although see Eye imaging in section 3.1.7 for an example of how this has the potential to change). However, a subset of structured, coded, individual-level data about NHS-funded (but not privately funded) dental examinations and procedures and eye tests is provided as a regular data feed to central, national NHS organisations in each of the UK's four nations.[62] These data are mainly provided for financial management (including payment for services), monitoring and planning of the services provided, and so do not contain detail on clinical findings during the examinations, procedures or tests. The exclusion of privately funded activities from these national data collections means that the coverage of eye tests and dental procedures conducted in community settings is variably incomplete, depending on country-specific arrangements for how these are provided and funded.[63] However, the data are potentially useful for wider purposes, for example in research studies, especially if linked to other health data sources, to support the generation of novel insights.

62 In England the relevant organisation is the NHS Business Services Authority, an arm's length body of the Department of Health and Social Care. See https://www.nhsbsa.nhs.uk/ for more information. In Scotland, Wales and Northern Ireland, the relevant organisations are, respectively: Public Health Scotland (https://publichealthscotland.scot/our-areas-of-work/primary-care/), NHS Wales Shared Services Partnership (https://nwssp.nhs.wales/), Northern Ireland Health and Social Care Business Services Organisation (https://bso.hscni.net/directorates/operations/family-practitioner-services/).

63 All health datasets that collate data on NHS-funded care exclude privately funded activity. However, the proportion of unscheduled healthcare that is funded privately in the UK is extremely low, while it is somewhat higher for certain scheduled activities. Currently, regular NHS eye tests are free for everyone in Scotland (https://www.nhsinform.scot/care-support-and-rights/nhs-services/eyecare/nhs-community-eyecare/), while in England, Wales and Northern Ireland they are free for certain groups based on age, income and medical history (https://www.nhs.uk/nhs-services/opticians/free-nhs-eye-tests-and-optical-vouchers/; https://www.gov.wales/get-help-nhs-eye-care-costs; https://www.nidirect.gov.uk/articles/eye-care). NHS dental examinations are free for everyone in Scotland but for most people treatments are not (https://www.nhsinform.scot/care-support-and-rights/nhs-services/dental/receiving-nhs-dental-treatment-in-scotland/#dental-treatment-costs); in England, Wales and Northern Ireland there are charges for NHS dental examinations and treatments, which are provided by dental practices that offer a mix of NHS and private services (https://www.kingsfund.org.uk/insight-and-analysis/long-reads/dentistry-england-explained; https://www.gov.wales/nhs-dental-charges-and-exemptions; https://www.nidirect.gov.uk/articles/health-service-dental-charges-and-treatments; https://www.dentalhealth.org/paying-for-dental-treatment-in-the-united-kingdom). Access to and uptake of NHS dental services remains challenging across the UK (https://www.mydentist.co.uk/docs/default-source/default-document-library/gbohr/the-great-british-oral-health-report-2021.pdf).

National-scale data on a range of other community health services

In England, publicly funded community care providers submit data monthly to NHS England for the Community Services Dataset. This includes national patient-level data about publicly funded community health services for children, young people and adults, provided in settings such as health centre, Sure Start centres, day care facilities, schools, community centres, mobile facilities or patients' homes. The data include information on personal and demographic details, social and personal circumstances, breastfeeding and nutrition, care event and screening activity diagnoses, including long-term conditions and disability assessments.[64] Similar national data are collected in the devolved administrations.[65]

3.1.4 Data from hospitals

Detailed information about the wide range of conditions treated

Hospital-based doctors, nurses, therapists and other healthcare professionals in hospitals across the UK record information in hospital computer systems about patients cared for in various parts of the hospital, including accident and emergency departments, specialist outpatient clinics, operating theatres, day case and inpatient wards and intensive care units. The information recorded includes documentation of symptoms, signs, observations, measurements, diagnoses, treatments, operations and other procedures, and clinical correspondence (for example letters summarising specialist outpatient consultations or 'discharge summaries' of stays in hospital for future reference and for sharing with other parts of the NHS, especially general practices). By contrast with general practice, information on patients staying in hospital will often be generated and recorded daily, many times per day or even continuously for seriously ill patients in intensive care.

64 See https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services/data-set-catalogue/community-services-data-set-csds.

65 E.g. see https://publichealthscotland.scot/services/data-management/data-management-in-primary-social-and-community-care/overview/who-we-are/; https://www.datadictionary.wales.nhs.uk/.

Electronic patient record systems were adopted later and are more fragmented than in general practice

The adoption of electronic patient record (EPR) systems to replace paper records in NHS hospitals across the UK occurred much later than in general practices and has had a chequered history. Attempts during the 2000s to roll out a centrally managed IT system for the NHS across England were fraught with challenges. After spending billions and failing to deliver against many of its ambitious goals (albeit with some notable successes[66]), this so-called 'NHS National Programme for IT' was discontinued in 2011. It was replaced within a few years by new programmes of investment (the Global Digital Exemplar and Local Health and Care Records Programmes). These took a phased, regional approach, which focused on introducing EPR systems in hospitals, improving interoperability between primary, secondary and social care systems, and training NHS leaders in clinical informatics.

There has been progress as a result, in that almost all hospitals (around 90%) across England now have an EPR system. However, the many commercial hospital EPR systems (possibly >40),[67] compared with the very small number for primary care, bring greater interoperability challenges (both regionally and nationally), since such challenges generally increase in line with the number of different systems and supplier organisations. And, while there are examples of data integration enabling more joined-up care for some patients in some areas, the goal of interoperability at regional level across primary care, secondary care, social care and other systems in each of England's 42 integrated care systems (ICSs – each covering a population of between 0.5 and 3.5 million) is still a long way off.[68] Interoperability could be improved by the application of information standards to IT suppliers in the health and social care system as part of the new government's forthcoming Digital Information and Smart Data Bill.[69]

The devolved nations, Scotland, Wales and Northern Ireland (population 5.5, 3.1 and 1.9 million respectively), each cover a population roughly the size of one to two large English ICSs. Each is moving towards a position where most or all hospitals will use the same EPR system,[70] with an ambition that this will facilitate interoperability with primary care and social care EPR systems. However, as in most parts of England, these systems are still being rolled out and are some way from being fully integrated and functional.

66 E.g., supporting the building of the NHS England Spine (https://digital.nhs.uk/services/spine), which securely holds all NHS England patients' demographic information (Personal Demographic Service) and now allows information to be shared securely through national services such as the Electronic Prescription Service, the Personal Demographics Service, the Summary Care Record and the e-Referral Service.

67 We were unable to find clear information in the public domain or to obtain accurate estimates from NHS England. However, in 2022, Digital Health Intelligence assessed the digital maturity of 132 NHS England acute trusts, reporting that eight different commercial electronic patient record systems were deployed in the 32 most digitally mature trusts, but providing no information on the systems used in the remaining 100 trusts surveyed (see https://digitalhealthintelligence.net/digital-maturity-acute-nhs-snapshot-report/. Simon Bolton, interim CEO at NHS Digital and chief information officer at NHS England from mid-2021 to early 2023, reported in November 2022 that there were 40-60 different electronic patient record systems across hospital trusts in England (see https://www.computerweekly.com/news/252527290/Commoditisation-of-NHS-tech-is-a-problem-says-NHS-Digital-interim-CEO/).

68 In England, integrated care systems (ICSs – see https://www.england.nhs.uk/integratedcare/ and https://www.kingsfund.org.uk/insight-and-analysis/long-reads/integrated-care-systems-explained) were introduced in July 2022, replacing Clinical Commissioning Groups, which themselves replaced Primary Care Trusts in April 2013.

69 See https://assets.publishing.service.gov.uk/media/6697f5c10808eaf43b50d18e/The_King_s_Speech_2024_background_briefing_notes.pdf.

70 In Scotland, hospitals in 12 of 14 health boards now use a version of TRAKCare, provided by the company Intersystems (https://www.intersystems.com/uk/success-stories/unifying-healthcare-in-scotland/, https://www.publichealthscotland.scot/publications/acute-hospital-activity-and-nhs-beds-information-quarterly/acute-hospital-activity-and-nhs-beds-information-quarterly-quarter-ending-30-june-2023/data-quality/); in Wales, rather than contracting with a commercial supplier for a hospital-based EPR, integrated electronic data capture systems are being iteratively developed and integrated within the single Welsh Clinical Portal accessible to all relevant health and social care staff across Wales (https://dhcw.nhs.wales/ig/ig-documents/imtp-23-26/); in Northern Ireland, a single electronic patient record system provided by the company EPIC is being rolled out across hospital trusts in Northern Ireland from late 2023 to 2025 (https://www.health-ni.gov.uk/news/date-set-patient-record-revolution).

Well-established national data collections about episodes of care in NHS hospitals

EPR systems in secondary (hospital) care are important for the delivery of modern hospital patient care. However, national systems for the collection from hospitals of sets of structured, coded data to monitor activity, health trends and costs existed in all four nations of the UK years before the implementation of EPR systems across UK hospitals. These include hospital episode statistics (HES) in England, available from 1997 onwards, Scottish Morbidity Records (SMR) in Scotland, available from 1981 onwards, patient episode data for Wales (PEDW) in Wales, available from 1998 onwards, and hospital admissions and discharges data for Northern Ireland, available from 2000 onwards. These national collections summarise key features of hospital episodes of care for individual patients cared for as inpatients, day case patients and outpatients, as well as in emergency departments, critical care, psychiatric care and maternity care. The data include administrative details such as dates of appointments, admission and/or discharge as well as codes for diagnoses and operative procedures. The records can be linked at person level to other sources of health-relevant data for the entire population of each country or for defined subsets of the population (for example a regional sub-population, or the participants in a research cohort such as UK Biobank or a clinical trial such as RECOVERY).

National hospital data have great value but important limitations

National hospital data do not include the depth and granularity available within hospital EPR systems (for example extensive unstructured narrative 'free text' from ward activity, clinical correspondence and discharge summaries). But their national scale, linkability, relevance to a broad range of health conditions involving hospital care, and years of coverage (extending backwards many years as well as forwards over time) make them hugely valuable for research and analysis (see section 1.1). There are already plenty of examples of their use for a wide range of beneficial purposes, a few of which are shown in Box 3.3. However, as with the other sources of data from the healthcare system reviewed here, more streamlined, extensive and broader uses would substantially extend and magnify those benefits.

Box 3.3 Examples of uses of hospital episode statistics (HES) data

Monitoring hospital activity

HES data are used by NHS England to produce regular reports on activity in English hospitals by age, speciality and admission type. For example, the graphs on the right show numbers of hospital inpatient episodes and admissions of different types for all specialities (upper graph) and for cardiac surgery (lower graph) for the period June 2022 to May 2024:[71]

Understanding inequalities in healthcare use

The Office for National Statistics uses records from the latest Census (2021) linked via NHS number to NHS England's Emergency Care Data Set (ECDS) and Hospital Episode Statistics (HES) to study patterns in emergency care attendance in England. The most recent report showed that A&E department attendance increased with increasing levels of deprivation, after adjusting for age, sex, and ethnicity. It also showed that this relationship was partly – but not completely – explained by poorer health in more deprived people, who may also have poorer access to primary care.[72]

Following the health of people in longitudinal research cohorts – the example of UK Biobank

The main method used to follow the health of the 500,000 participants in the population-based cohort, UK Biobank, is through linking to data about the participants in national health databases, including hospital episode statistics and equivalent datasets across England, Scotland and Wales, where the participants are based. These linked health data have been used in thousands of research studies.

One example is a study of the relationship between objectively assessed physical activity, measured in 90,000 participants who wore wrist-watch-like accelerometer devices that recorded their physical activity continuously for seven days, and the later development of cardiovascular disease, ascertained from hospital episode statistics. The research team found that, compared with the least active people, the most active people had less than half the risk of stroke and heart attack over the next five years. This protective effect of physical activity was much stronger than previously thought from the findings of studies that had assessed physical activity through less accurate self-report questionnaires.[73]

71 See https://digital.nhs.uk/data-and-information/publications/statistical/provisional-monthly-hospital-episode-statistics-for-admitted-patient-care-outpatient-and-accident-and-emergency-data/april-2024---may-2024.

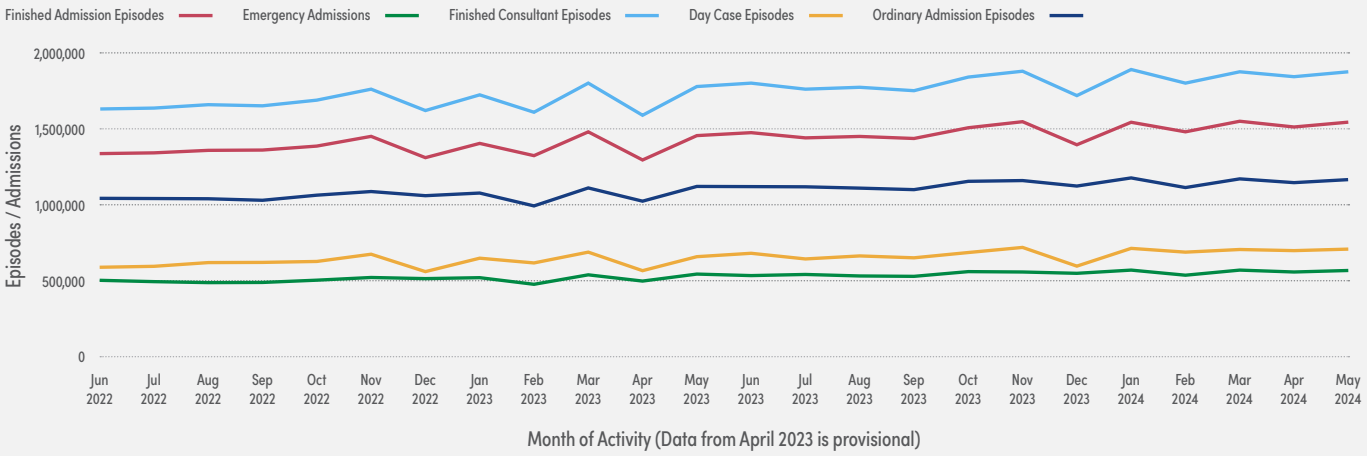
72 See https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthcaresystem/articles/inequalitiesinaccidentandemergencydepartmentattendanceengland/march2021tomarch2022.

73 See Ramakrishnan R et al. *Accelerometer measured physical activity and the incidence of cardiovascular disease: Evidence from the UK Biobank cohort study.* PLoS Medicine 2021 (https://journals.plos.org/plosmedicie/article?id=10.1371/journal.pmed.1003487).

Provisional Monthly HES data for Admitted Patient Care by Treatment Specialty

Treatment specialty name: All

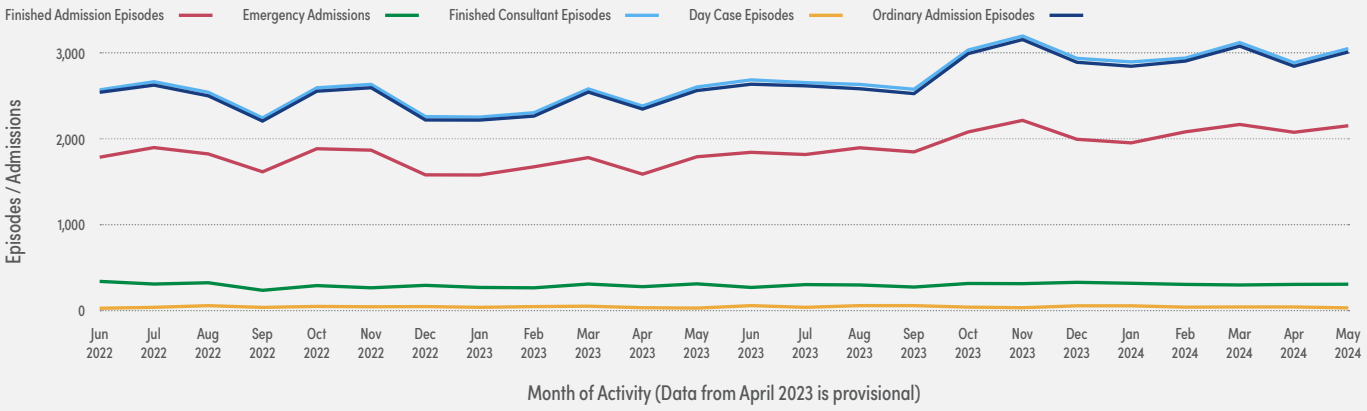
Inpatient Monthly Activity by episode / admission type



Provisional Monthly HES data for Admitted Patient Care by Treatment Specialty

Treatment specialty name: 172 Cardiac Surgery Service

Inpatient Monthly Activity by episode / admission type



Two key limitations across all four nations of these national collections of hospital episodes of care data are:

- the low percentage of outpatient episodes that are assigned diagnosis and/or procedure codes (3–4%);[74]
- the lag behind real time of the admitted patient episodes data, which are not submitted until after completion of an episode of care (which may be short or prolonged) and take additional time to curate before and after submission.[75]

These limitations are important. The paucity of coded diagnostic and procedural information in outpatient data limits their broader uses. For example, because the patient's diagnosis is generally not captured in these national data, the data cannot readily be used alongside other national sources to identify and follow patients with a particular health condition (such as heart failure, asthma or dementia). Of note, specialist hospital outpatient diagnoses and procedures may be coded in general practice records, once correspondence about the outpatient consultations has been received by each relevant patient's general practice. However, the completeness and accuracy of this process across the range of conditions dealt with in specialist outpatient settings is uncertain. And it does not generate information in real time.

The lag of weeks to months behind real time for the assembly and availability of national hospital admissions data is not a problem for all the potentially beneficial uses, but it does seriously limit some critically important national-scale uses. For example, even with focused efforts to streamline national data collection at the height of the COVID-19 pandemic,[76] national hospital admissions data could not provide timely information about the diagnoses of patients with admissions of prolonged duration (generally the sickest patients). This meant that data from some of these patients were omitted from analyses and reports based on national hospital admitted episodes data. For example, in the early months of the COVID-19 vaccination programme, clinicians observed serious (but rare) adverse effects that were thought to be related to one or more of the vaccines being used. The time lag in data on hospital diagnoses of potential vaccine adverse effects meant that analyses of the benefits versus the risks of different vaccines could not be generated as rapidly as would have been ideal to inform key decisions on population vaccination policy.

74 Diagnosis (International Classification of Diseases [ICD-10]) and procedure (international classification of interventions and procedures OPCS4) classification codes (see https://classbrowser.nhs.uk/#/) for the national data collection submissions from hospitals are routinely applied for hospital admission (including day case) episodes of care but only for a minority (3–4%) of hospital outpatient ones.

75 In most hospitals, coding of admitted episode of care is done after the end of a hospital consultant episode of care or after discharge, when trained hospital coding clerks apply these codes, usually based on a semi-structured discharge summary completed by a doctor (or, if this is unavailable, using other information from the electronic patient record [EPR]). The roll-out and optimisation of EPRs across hospitals in the UK, together with developments in automated assignment of codes from free text using natural language processing, brings the potential for increasingly automated real-time coding of hospital care using internationally recognised **point of care clinical classification schemes**, in particular SNOMED-CT (the main coding schema used in general practice). Few UK hospitals currently implement such realtime coding of clinical encounters.

76 E.g. through the Chief Scientific Adviser's Data and Connectivity National Core Study (https://www.hdruk.ac.uk/covid-19-data-and-connectivity/).

Real-time data on hospital diagnoses, linked to real-time information from other data sources,[77] are a crucial part of the national-scale capability needed to track newly emerging health conditions and other health threats, to assess the impact of these on already established diseases, and to monitor the adverse effects of new medicines and other healthcare products. Such capability is also needed for a range of other purposes, for example real-time national monitoring of the safety of all new vaccines, drugs and devices.

3.1.5 Data on prescribed and dispensed medicines

Medicines data come from a range of sources

Medicines data are of substantial interest to healthcare planners, researchers, medicines regulatory and approval bodies, policymakers, patients, the wider public and others, and are relevant across a broad spectrum of health conditions. They are important economically: the UK's total expenditure on medicines in 2022 was £36.7 billion – over 12% of healthcare costs in that year.[78] Given the potential for harm as well as benefit from medicines, as highlighted in the Cumberlege Independent Medicines and Medical Devices Safety Review in 2020,[79] high-quality, comprehensive data on medicines are also critical for monitoring safety.

Several sources of medicines prescribing and dispensing data from healthcare settings are collected and collated nationally by NHS organisations in each of the four nations. These different sources of medicines data can be linked at individual patient level to each other and to other health data sources. Recent years have seen substantial progress in each of the four UK nations in the national collation and curation of some of these datasets, and their linkage to other health data sources (including data from primary care and hospitals) at whole-population scale. These developments have started to demonstrate the potential for large-scale analyses of geographical variation, trends over time, adherence to guidelines, costs, and both positive and negative impacts on health outcomes of the prescribing and dispensing of a wide range of medicines for many different health conditions. A couple of examples motivated by the COVID-19 pandemic illustrate this (Box 3.4). However, gaps and substantial untapped potential remain.

77 such as general practice data for diagnoses made in general practice, or national infectious disease diagnostic testing data where relevant.

78 2022 data from the Office for National Statistics, see: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthcaresystem/bulletins/ukhealthaccounts/2022and2023.

79 See https://immdsreview.org.uk/Report.html.

Box 3.4 Benefits of linking data on the prescribing and dispensing of medicines to other health data sources at whole-population scale

Analysing medicines data from 17 million people to understand the risk of severe COVID-19 in people with immune-mediated inflammatory diseases[80]

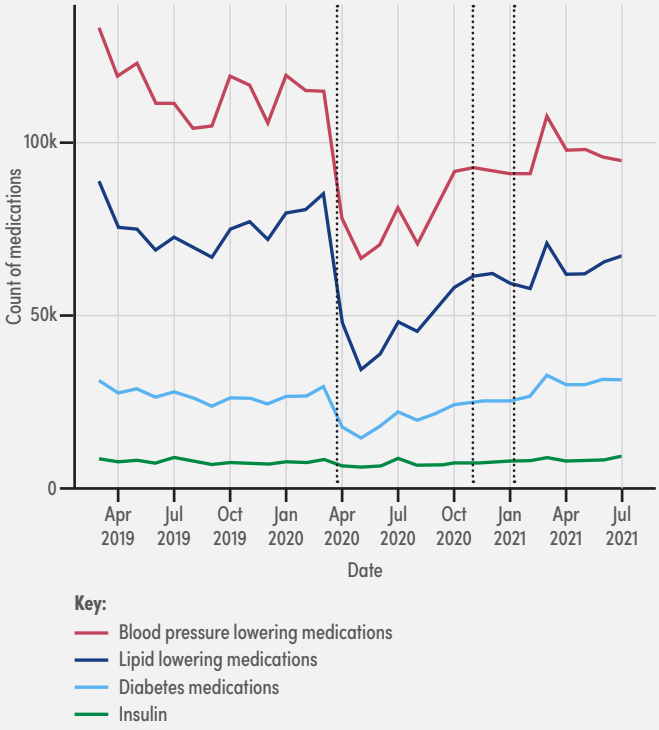
Researchers analysed hospital medicines prescription data linked to routinely collected general practice data, hospital admission and death data from 17.7 million adults in England, of whom 1.2 million had immune-mediated inflammatory diseases. They found a higher rate of hospital admissions and deaths due to COVID-19 in people with immune-mediated inflammatory diseases. However, the rate of severe COVID-19 was not higher in those on the most targeted immune-modifying drugs for immune-mediated inflammatory diseases compared with those on standard immune-modifying drugs.

Insights from data on medicines to prevent cardiovascular diseases from a population of over 60 million people across England, Scotland and Wales[81]

Researchers analysed data from England, Scotland and Wales on medicines for the prevention of cardiovascular diseases (including heart attack and stroke), prescribed and dispensed in the community, linked with other sources of health data. They found that use of blood pressure-lowering medicines, lipid-lowering medicines and diabetes medicines (but not insulin) fell markedly during 2020–2021 compared with before the pandemic (see graph).

They reported that almost half a million fewer people than expected started antihypertensive treatment in England, Scotland and Wales from March 2020–May 2021.

They estimated that the resulting under-treatment of raised blood pressure would have caused over 13,500 additional vascular events, including over 2,200 additional heart attacks and over 3,400 additional strokes.



80 McKenna B et al. *Risk of severe COVID-19 outcomes associated with immune-mediated inflammatory diseases and immune-modifying therapies: a nationwide cohort study in the OpenSAFELY platform.* Lancet Rheumatology 2022 (https://pubmed.ncbi.nlm.nih.gov/35698725/).

81 See Dale C et al. *The impact of the COVID-19 pandemic on cardiovascular disease prevention and management.* Nature Medicine 2023 (https://www.nature.com/articles/s41591-022-02158-7).

The two main sources of medicines data arising from day-to-day activity in the health and care system are community prescribing and dispensing, and hospital prescribing and administration.

Community prescribing and dispensing data: national-scale data in all four UK nations

Information on prescriptions in general practice is recorded in general practice electronic records (see section 3.1.2) and provides a rich source of information about NHS prescriptions in primary care across the UK. Complementing the general practice data are data from over 13,000 community-based pharmacies across the UK.[82] Community pharmacies incorporate prescription information into computer systems that hold data on the medicines and other items they dispense. The data in these systems provides useful information not only on medicines that are prescribed but also on those that are dispensed.[83] Much of the data from community pharmacy systems are collected and collated nationally in structured, coded and linkable format by national NHS organisations in each of the four UK nations, primarily to enable service planning, monitoring and financial management.[84]

In Scotland, national community prescribing and dispensing data have been curated in the Prescribing Information System[85] and made securely available for healthcare planning and research for well over a decade, with linkages to a range of other health data. Developments over the last few years have seen similar national datasets starting to become similarly accessible in England, Wales and Northern Ireland. Although driven by the need for data to monitor and understand the impact of the COVID-19 pandemic, these developments have far wider potential benefits because medicines are prescribed for such a broad range of health conditions. England's very large population size makes the availability of these data particularly noteworthy. Prescribing and dispensing data from all community dispensing outlets across England are now provided regularly by the NHS Business Services Authority to NHS Digital (now part of NHS England). These data can be linked to other sources of health data (including from primary and secondary care) at national scale.

82 Community pharmacies are usually independent businesses contracted by the NHS to provide certain services for local populations. There are over 13,000 community pharmacies UK-wide (around 11,000 in England, 1200 in Scotland, 700 in Wales and 500 in Northern Ireland). See https://www.kingsfund.org.uk/insight-and-analysis/long-reads/community-pharmacy-explained; https://www.gov.scot/policies/primary-care-services/pharmacy/; https://www.gov.wales/community-pharmacy-services-april-2022-march-2023-html; https://www.health-ni.gov.uk/news/publication-general-pharmaceutical-services-northern-ireland-annual-statistics-202223.

83 For various reasons, not all prescribed medicines are either dispensed or taken. The fact of a medicine being dispensed gets a step closer to (but does not guarantee) that medicine being taken as prescribed.

84 In England the NHS Business Services Authority (https://www.nhsbsa.nhs.uk/), in Scotland, Public Health Scotland (https://publichealthscotland.scot/our-areas-of-work/primary-care/), in Wales, NHS Wales Shared Services Partnership (https://nwssp.nhs.wales/), in Northern Ireland the NI Health and Social Care Business Services Organisation (https://bso.hscni.net/directorates/operations/family-practitioner-services/).

85 https://find.researchdata.scot/dataset/22e3943e-edb5-44a1-9e4e-22b0f7a31767.

Hospital prescribing and administration data: systems do not yet scale nationally

In an increasing number of hospitals (but not yet all) across the UK, information about the medicines prescribed and given to patients in hospital is entered into electronic prescribing medication administration (EPMA) computer systems, replacing paper drug charts. These systems may be integrated within the hospital's main EPR system or sourced from one of a range of EPMA system suppliers[86] and deployed independently of the hospital's main EPR. The capture of hospital medicines information in electronic form is a step towards its further use for wider benefit, particularly when linked to other data sources, such as data about health conditions or tests before and after certain medicines are administered.

The collection of national data from EPMA systems, linkable to other national health data sources, started in both England and Scotland during the COVID-19 pandemic. In England, NHS England (then NHS Digital) established a daily collection of data from one of the most widely used EPMA systems, which covered 10–15% of hospital NHS trusts. These data were made securely available for COVID-related analysis and research, with linkage to other data sources. However, the limited geographic coverage has restricted the usefulness and usability of these data. The collection of these data for COVID-related purposes was paused in August 2023,[87] but from January 2025 will be replaced by a national collection, incorporating data from additional EPMA systems and encompassing a wide range of purposes beyond COVID-19.[88] This will be a major step forward.

EPMA systems are also being successfully rolled out across Scotland, and currently cover around two thirds of the 5.5 million population. Public Health Scotland has established a regularly updated Scottish national hospital EPMA dataset through regular automatic extraction of data from these local systems. The dataset can be linked to other sources of Scotland-wide health data and made securely available for a range of research and other analyses.[89] In Northern Ireland, an EPMA system is included as part of the new EPIC hospital EPR system that is being rolled out across all hospital trusts in the country (see section 3.1.4).

86 See https://digital.nhs.uk/services/digital-and-interoperable-medicines/resources-for-health-and-care-services/list-of-epma-suppliers.

87 See https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/electronic-prescribing-and-administration-epma-data-in-secondary-care.

88 See https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/secondary-care-electronic-prescribing-and-medicine-administration-epma-data-collection.

89 See https://ijpds.org/article/view/2182.

Other sources of medicines data add complexity

Unfortunately, the two main community- and hospital-based medicines sources described in the previous sections are not the only sources of medicines data. For example, systemic anti-cancer treatment (SACT, more widely referred to as 'cancer chemotherapy') data have separate prescribing, administration and national data collection processes in all four nations of the UK.[90] In addition, separate systems are used in England for the approval and management of many high-cost drugs,[91] which are therefore not included in the main community- and hospital-based medicines sources. In an attempt to address this gap, a national high-cost drugs dataset was generated during the COVID-19 pandemic by collating payment submissions data on all high-cost drugs from hospitals across England.[92] This dataset was rapidly used to better understand the benefits, risks and uses of high-cost drugs during the pandemic (see the first example in Box 3.4). Unfortunately, rather than being established as a regularly updated national data collection for England, the collection of this national high-cost drugs dataset has to date remained as a one-off exercise, but it has at least demonstrated a process for its collection and the benefits of doing so.

National vaccination data systems accelerated during the pandemic

General practice records are currently the main source of information on vaccines administered to adults and children. The completeness and accuracy of these data rely both on the reliable transfer into general practice computer systems of data about vaccines delivered in various non-practice settings (for example schools or pharmacies) and on the correct coding of these data in the general practice systems.

However, in all four nations of the UK, the COVID-19 pandemic led to the accelerated development of national vaccination data systems, enabling the direct entry of data from multiple sites administering COVID-19 vaccines, including hospitals, general practices, pharmacies, mobile vaccination units, and mass immunisation sites. These systems were designed both to collect and hold data nationally as well as to enable rapid, efficient transfer of information on vaccinations into general practice computer systems. In England and Scotland, the systems have expanded to incorporate data on vaccinations given to protect against some non-COVID-19 infections (for example influenza). These national vaccination systems have been essential in supporting streamlined, real-time monitoring of COVID-19 vaccine uptake across the UK by national public health organisations.[93] And the secure availability of these data at patient level, linked with health data from primary care, secondary care and death registers has enabled important research across the whole UK population into the drivers and consequences of under-vaccination against COVID-19.[94]

90 E.g. see https://digital.nhs.uk/ndrs/data/data-sets/sact (England); https://www.nature.com/articles/s41416-021-01262-8 (Scotland).

91 Mainly, but not exclusively, a system called Blueteq (https://www.blueteq.com/commhcd.html; https://www.england.nhs.uk/publication/nhs-england-drugs-list/).

92 This work was coordinated by the Oxford-based OpenSAFELY team in partnership with NHS England, with more details here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120928/.

93 In England, the National Immunisation Management System includes COVID-19 and influenza vaccine data (https://www.sciencedirect.com/science/article/pii/S138650562200288X); in Scotland, the Turas Vaccination Management Tool is used in vaccine uptake monitoring for influenza, COVID-19, pneumococcus and shingles and is configured for additional vaccines (https://scotland.shinyapps.io/phs-vaccination-surveillance/); the Welsh Immunisation System and Northern Ireland Vaccine Management System both include COVID-19 vaccine data (https://www2.nphs.wales.nhs.uk/CommunitySurveillanceDocs.nsf; https://www.health-ni.gov.uk/vaccines-management-system-response-covid-19).

94 https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(23)02467-4/fulltext.

3.1.6 Laboratory data

Huge numbers of tests and test results

NHS clinical and diagnostic laboratories, mainly based in hospitals, analyse samples (for example of blood or urine) taken from patients for a wide range of tests to help inform diagnosis, prognosis and treatment choices. These laboratories handle test requests from both general practice and hospital settings and generate very large numbers of data items. For example, across the NHS in England, pathology laboratories carry out over a billion tests annually at a cost to the NHS of around £2.5 billion, generating an estimated 100 billion data items.[95] Laboratory test results account for a third of all coded data items in general practice records; and the volume of data generated from hospital tests is around 100 times higher than those generated from general practices.

Complex network of laboratory computer systems

Laboratories use computer systems called laboratory information management systems (LIMS) to handle data on test requests and results. Single national LIMS systems are already in place or being implemented in Scotland, Wales and Northern Ireland to facilitate the sharing of laboratory data in healthcare settings across the country. In England, multiple NHS laboratories use many different LIMS. Large numbers of so-called 'middleware' software systems (sometimes referred to as 'interoperability patches') are needed to enable sharing of information between laboratories and other computer systems (for example hospital EPR systems). National pathology messaging systems are also used across the UK for the exchange of requests and results between laboratories and EPR systems in primary care,[96] and between laboratories.[97] In England in particular, the flows of data between different systems to allow sharing of laboratory data are highly complex (Figure 3.1). While separate components of this complex system may work well, we heard from laboratory data experts that it does not facilitate data sharing and accessibility at national scale, which remains highly desirable.

95 https://digital.nhs.uk/services/pathology-standards-and-implementation.

96 E.g. in England: https://digital.nhs.uk/developer/api-catalogue/pathology-messaging-fhir; in Wales https://dhcw.nhs.wales/systems-and-services/secondary-care/welsh-laboratory-information-management-system/; in Scotland, laboratory test information is transmitted to and stored in a system called SCIStore, which is implemented separately across Scotland's 15 health boards but supports sharing of information between health boards and with general practice systems (https://www.sci.scot.nhs.uk/products/store/store_main.htm); both Scotland and Northern Ireland are in the process of implementing national-scale LIMS systems (https://www.magentus.com/nhs-scotland-awards-national-laboratory-medicine-framework-to-magentus/?r=e; https://bso.hscni.net/directorates/digital-operations/nipims/laboratory-information-management-system-lims/).

97 Since 2007, the company X-Lab has provided a National Pathology Exchange (NPEx) across England, Scotland, Northern Ireland and Wales. Now known as Labgnostic, NPex allows laboratories to connect through a single hub, enabling test requests and pathology results to be sent digitally between labs. Prior to the pandemic, about 60% of UK laboratories used Labgnostic. During the pandemic the UK government awarded X-Lab a contract to connect additional laboratories to meet coronavirus testing demands. Labgnostic now services 95% of UK laboratories (https://x-labsystems.com/products/labgnostic/).

National standards for laboratory data needed

Not only do the many different laboratories use a wide range of different computer and software systems, they also record data from tests using many different data formats and coding systems. Over many years, clinical laboratory and data specialists working within and across the four nations have worked towards the adoption of national standards for recording laboratory test data. However, further development and implementation of national (UK-wide) standardised terminology (the language used to describe each test), measurement units and reference ranges across laboratories, together with the systems to translate existing laboratory data into such a standard, are still needed.[98] These developments will need consistent investment in long-term planning, leadership, and capacity in clinical informatics and bioinformatics. They will be much more challenging for complex tests with results that rely on free text narrative reporting rather than simple numeric measures. But focusing initial efforts on the more straightforward tests with simple numeric results could cover a lot of ground relatively quickly, since a relatively small number of these tests account for a very large proportion of all laboratory data.[99]

National, standardised, integrated systems for laboratory data across the UK: an aspirational goal

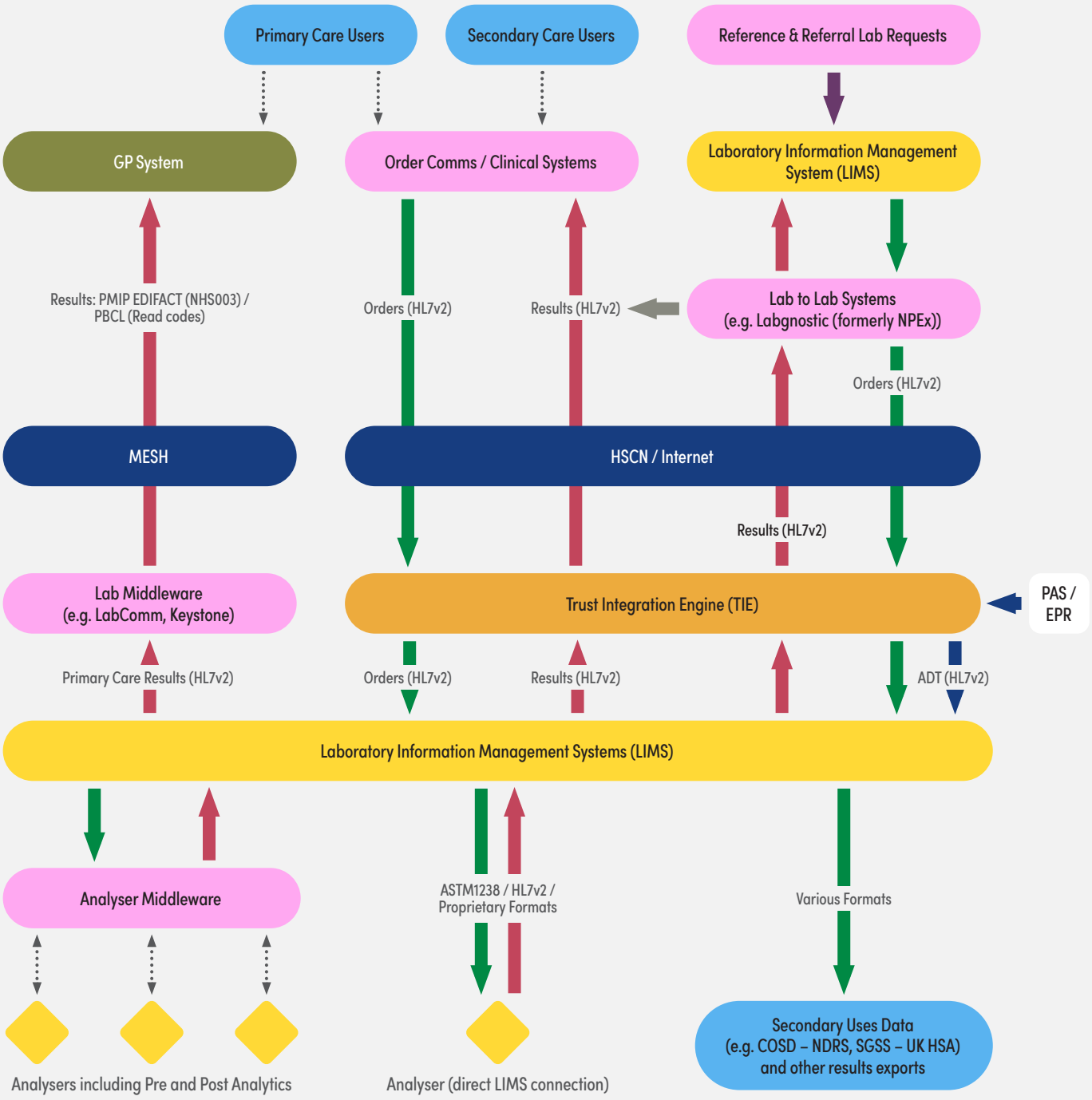
Pathology specialist leaders and laboratory data experts across the four nations of the UK remain passionate about the need for national, standardised, integrated systems for laboratory data. Achieving this goal will require the adoption of national data standards and improved computer system interoperability. The prize would be systems in each of the four UK nations that:

- allow clinicians across primary and secondary care settings to view test results from NHS laboratories outside their organisation or location, ensuring timely availability of key clinical information and avoiding unnecessary duplication of tests;
- enable monitoring and standardisation of laboratory practice, leading to a reduction of inappropriate test requests (reducing costs) and the harmonisation of reporting and reference ranges;
- transform national research and analysis capability through access to standardised national patient-level data linked to other national health data sources. This would unleash opportunities to better characterise health and disease, understand mechanisms of disease, and develop new approaches to diagnosis and treatment. For example, laboratory test results are important for accurately identifying patients with health conditions such as kidney and liver diseases, diabetes and arthritis. A national laboratory data resource would allow such patients to be far more accurately identified for inclusion in national analyses and research studies than is currently possible.

98 For an update on current status in England, see https://digital.nhs.uk/services/pathology-standards-and-implementation.

99 E.g., we learned from specialist NHS England staff that less than one hundred out of several thousand SNOMED-CT laboratory test codes account for over 90% of the total volume of laboratory test codes used in general practice computer systems.

Figure 3.1 The complex landscape of pathology systems across the NHS in England[100]



100 Figure provided by NHS England laboratory data information specialists, included here to demonstrate the complexity of the systems handling laboratory data.

National systems for some laboratory data demonstrate the art of the possible

While national systems for the standardisation and collation of data on most laboratory tests are some way from being a reality, a few national laboratory collection systems for test data exemplify what is possible (Box 3.5).

Box 3.5 Examples of national systems for laboratory data in the UK

National microbiology data systems

The Second-Generation Surveillance System (SGSS) is a national microbiology laboratory reporting and data collection system used in England to capture routine laboratory data on infectious diseases and antimicrobial resistance. It is managed by the UK Health Security Agency (UKHSA). Diagnostic laboratories are legally required to provide standardised data to the UKHSA within a week of specified micro-organisms causing certain communicable diseases being found in a human sample.[101] SGSS data are stored centrally within the UKHSA and may be securely shared with other organisations for public benefit. For example, during the COVID-19 pandemic, sharing of data between the UKHSA and NHS England enabled patient-level linkage of records on COVID-19 testing with other primary and secondary care health data sources at national scale. This allowed approved researchers to assess the impact of prior health risk factors and health conditions on COVID-19 occurrence and severity as well as the impact of COVID-19 on a wide range of health outcomes, providing vital information to guide healthcare and public health policy.

Although it has the major advantage of being a national system, the SGSS does have limitations. For example, data sharing is not mandatory for micro-organisms that are not specified in the Health Protection (Notification) Regulations (2010) (for example HIV is not specified). And only positive (but not negative) test results are shared, which means that meaningful comparisons

101 See https://assets.publishing.service.gov.uk/media/66e2e0ba0d913026165c3d77/UKHSA_Laboratory_reporting_guidelines_May_2023.pdf and https://www.gov.uk/guidance/notifiable-diseases-and-causative-organisms-how-to-report.

cannot be made between people who have had a positive versus a negative test.

Similar systems for national microbiology laboratory reporting of communicable diseases exist in Scotland, Wales and Northern Ireland.[102]

National genomic sequencing data in England

In England, genomic testing is conducted through NHS England's Genomic Medicine Service.

NHS genomic tests that are not based on genomic sequencing but targeted at specific genes, panels of genes or regions of the genome are conducted by a network of seven Genomic Laboratory Hubs, established in 2018. (There are similar laboratories in Scotland, Wales and Northern Ireland[103]). Each hub coordinates genomic testing services for a particular part of the country, following a single National Genomic Test Directory.[104]

The IT infrastructure underlying these regional genomic testing services remains 'clunky' (for example, paper forms are still used to request genomic tests in many English hospitals) and not yet able to support a national repository or data access system for genomic test results.

By contrast, whole genome sequencing and associated data analysis is coordinated by Genomics England (GEL) as a national service for an increasing set of specific indications (for example cancer, rare diseases).[105] GEL returns diagnostic interpretation results to the Genomic Laboratory Hubs. The rationale for a national service is compelling: genomic sequence data are complex, high volume (a single person's whole genome sequence comprises billions of data points), and need specialised systems and processes for their management and analysis. Provided patients have given their consent, Genomics England also stores the genomic sequence data in a secure national database, the National Genomic Research Library, where it can be linked to other national sources of health data and accessed securely for approved research studies.[106]

National laboratory data in Wales

Data from Wales's single national laboratory information system feed into the national Welsh Results Reporting Service (WRRS), allowing healthcare professionals across Wales to access, enter and view laboratory test requests and results. Through a partnership between data science experts at Swansea University and NHS Wales, work is underway to standardise and curate these data to create a Welsh national laboratory data resource, linkable to other sources of health data within the Welsh SAIL databank.[107]

102 See https://publichealthscotland.scot/services/national-data-catalogue/national-datasets/a-to-z-of-datasets/electronic-communication-of-surveillance-scotland-ecoss/; https://www2.nphs.wales.nhs.uk/CommunitySurveillanceDocs.nsf; https://www.publichealth.hscni.net/directorate-public-health/health-protection/surveillance-data.

103 For Scotland, see https://www.nss.nhs.scot/specialist-healthcare/specialist-services/genetic-and-molecular-pathology-laboratories/; for Wales, see https://medicalgenomicswales.co.uk/index.php/download-services; in Northern Ireland, the Belfast Health & Social Care Trust Regional Molecular Diagnostics Service delivers genetic and molecular diagnostic services to all five health trusts in Northern Ireland.

104 https://www.england.nhs.uk/genomics/nhs-genomic-med-service/.

105 https://www.genomicsengland.co.uk/initiatives.

106 See https://www.genomicseducation.hee.nhs.uk/supporting-the-nhs-genomic-medicine-service/national-genomic-research-library-information-for-clinicians/. By the end of 2022, the National Genomic Research Library held information on 135,000 genomes, with 65 petabytes of genomic and clinical data, and was providing secure access to data for >1,700 approved and registered researchers (https://www.genomicsengland.co.uk/pages/annual-report-2022/).

107 See https://popdatasci.swan.ac.uk/wp-content/uploads/2022/04/Data-Explained-The-Welsh-Results-Reports-Service-WRRS-Data.pdf.

3.1.7 Imaging data

Different sources and types of imaging data share similar challenges of data and system complexity

Imaging data is generated across many different parts of the NHS. These include:

- images from X-rays, computed tomography (CT), magnetic resonance (MRI), ultrasound and other types of scans, mainly conducted in hospital radiology departments;
- images of the retina at the back of the eye generated from scans done as part of eye tests in the community or in specialist ophthalmology hospital settings;
- images created from tissue specimens (for example from biopsies or surgical resections) examined in specialist hospital pathology departments.

Rapid developments in technology over the last few decades mean that images of all types are now increasingly created, stored, viewed and analysed digitally. This can and should bring major improvements across the NHS in two broad areas:

1. the efficiency of sharing of images between different parts of the NHS to benefit direct patient care;
2. the collection or integration of large sets of images that can be linked to other sources of health data to enable innovative research and analysis. This includes the development of tools including AI for rapid, efficient and accurate image processing and analysis. Such tools have the potential to relieve pressure on over-stretched NHS services as well as to support research to better understand health and disease.

Imaging data raise several specific challenges, most of them shared across imaging data types and sources. These arise from the complex, unstructured nature and higher volume of imaging data compared with structured, coded data from other sources, such as national hospital episode statistics or general practice records. The challenges include those of storage, transfer, standardisation of imaging data formats, robust de-identification and security protocols, analysis, and linkage to other health data sources. As with many other health data sources, poor interoperability across the many and varied computer systems for handling NHS imaging data is also a big issue.

Radiology imaging – some large-scale data resources in a fragmented landscape

Patients may have X-rays or scans in hospital radiology departments or community facilities to investigate and help diagnose the cause of their health problems. Almost all radiology imaging data across the UK are now acquired, viewed and stored electronically. Radiology information systems (RISs) are used to manage the administrative data about patients' imaging procedures, while Picture Archive and Communications Systems (PACSs) are used to store and view the images themselves. Data held within RISs are relatively simple and well-structured. The data from the images, held within PACSs, are complex, unstructured and of high volume (see Boxes 3.1 and 3.2) when considering the many different scans done each year across the UK, involving millions of people and generating billions of images.

Radiology services across the UK use many different RIS and PACS systems, which vary in their maturity, capability and interoperability. This variability affects the ease of sharing imaging data between hospitals for clinical care, as well as influencing data accessibility for broader benefits, such as planning of radiology services and research.

National imaging systems across the UK could achieve globally competitive healthcare and research

Two examples of national whole-population imaging data resources are NHS England's Diagnostic Imaging Dataset and the Scottish Medical Imaging resource (Box 3.6).

Box 3.6 National imaging data resources boosting health research and care

NHS England's Diagnostic Imaging Dataset

NHS Digital (now within NHS England) has collated a central, national dataset about diagnostic imaging carried out across the NHS in England since 2012 (the Diagnostic Imaging Dataset). The data are extracted from radiology information systems across the country and submitted monthly. For each imaging test, the Diagnostic Imaging Dataset captures patient demographics, the type and body site of the imaging, and the dates that the test was requested, performed and reported.[108] It does not include the images themselves or the results of radiologists' reports. However, it provides useful data for monitoring trends by imaging modality, geographic location and patient characteristics.[109] And linking these data to other health data at national scale allows the association of imaging procedures with subsequent health outcomes to be studied. Examples of such work in cancer and cardiovascular disease highlight the potential for further beneficial uses.[110]

108 See https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostic-imaging-dataset/. Note that the Diagnostic Imaging Dataset does not include information on radiology procedures not captured within hospital radiology information systems. These include breast screening, echocardiography, and many invasive radiology procedures.

109 E.g. see https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2024/03/Stat (istical-Release-21st-March-2024-PDF-305KB-1.pdf.

110 E.g. Fry A et al. *Linking the Diagnostic Imaging Dataset (DID) to cancer registration data - improving understanding of diagnostic imaging in lung and ovarian cancer.* Int J Pop Data Science 2017 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8362400/pdf/ijpds-01-131.pdf); Pearson C et al. *Establishing population-based surveillance of diagnostic timeliness using linked cancer registry and administrative data for patients with colorectal and lung cancer.* Cancer Epidemiology 2019 (https://www.sciencedirect.com/science/article/pii/S1877782119300517); Weir McCall JR et al. *National Trends in Coronary Artery Disease Imaging: Associations With Health Care Outcomes and Costs.* J Am Coll Cardiol Img 2023 (https://www.sciencedirect.com/science/article/abs/pii/S1936878X22006611?via%3Dihub).

Scotland's population-wide Scottish Medical Imaging data resource

Scotland was the first UK nation to introduce, from 2007–2009, a single national radiology Picture Archive and Communications Systems (PACS), leading to the rapid replacement of hard copy film with digital images. Following an upgrade in 2013, a modern replacement national PACS system is now being implemented.[111] A major advantage of this single national system is that a healthcare professional can view a patient's scans, regardless of where in Scotland the scanning occurred. This improves clinical decision-making and care, reduces unnecessary repeat scan requests, and allows radiologists across Scotland to view and report scans from any of Scotland's 14 health boards.[112]

This single national PACS has underpinned the development of the Scottish Medical Imaging data research resource: a copy of all NHS scans conducted across Scotland (for example CT, MRI, PET, X-rays, ultrasound) and associated data from the national PACS archive from 2008 onwards. Data from the first 10 years include around 2.5 billion images from 57 million imaging procedures, with a combined data volume of around 3 petabytes (see Box 3.2).[113] Pipelines have been established for these images to be de-identified, linked at patient level to other national sources of health data, and made securely available for approved research, including the development and testing of AI imaging tools. Many research studies using this resource are now underway to bring benefit for patients and the wider public.[114]

111 https://medical.sectra.com/news-press-releases/news-item/9C24FE58F8487231/.

112 In Scotland, better interoperability of RISs is also needed to support efficient sharing of radiology information and distribution of radiology reporting across the country.

113 This is a substantial volume of data, but very much smaller than other NHS data collections, e.g. the National Genomic Research Library, which holds 20 times the volume of data (65 petabytes).

114 See Baxter et al. *The Scottish Medical Imaging Archive: 57.3 Million Radiology Studies Linked to Their Medical Records.* Radiology: Artificial Intelligence 2024 https://pubs.rsna.org/doi/epdf/10.1148/ryai.220266.

Despite its now longstanding national PACS system, Scotland still has fragmented and poorly interoperable RISs. It aims to improve their interoperability through integration with the national PACS.[115] Wales has had single national RIS and PACS systems for several years,[116] while Northern Ireland is currently implementing a single combined national PACS and RIS system.[117] The key driver behind these is – rightly – the need to optimise clinical care through improved availability and sharing of imaging data between health professionals caring for patients. But the single system approach in Wales and Northern Ireland also heralds an opportunity to develop secure national imaging repositories, as in Scotland, to support research and innovation.

In England, around 20 imaging networks covering seven regions across the country were established in 2018.[118] One of their aims has been to promote the coordinated purchasing and commissioning of PACS and RIS systems, so that images can be more readily shared between hospital trusts across each network. This should improve clinical decision-making, aid the appropriate distribution of radiology reporting tasks, and better enable the deployment of new AI reporting tools for more efficient and automated workflows. However, although these networks have helped to align radiology services and personnel, sharing of images remains challenging. This is at least in part because attempts to move towards common RIS and PACS systems have been thwarted by trusts being locked into existing software supplier contracts of varying durations.

Based on their experience within and across imaging networks in England, radiology leaders have suggested the implementation of a platform (or platforms) that could pull images and related data from a variety of disparate PACS and RIS systems and hold a copy of them on a cloud-based server. Such a system could support viewing of images from multiple locations for clinical care and the sharing of reporting tasks. With appropriate implementation of de-identification protocols, drawing on experience in Scotland, it could also be used for the curation, analysis and secure access to or sharing of sets of images, linkable to other health data, for research. While implementation could be at either regional or national level, a national-scale endeavour would facilitate sharing of images not only within but also between regional networks. This could bring economies and efficiencies of scale. It would also enable effective care for patients in all geographic locations, including those whose care straddles different networks. A platform covering NHS imaging for the whole population of England would also create by far the largest whole-population imaging repository globally, bringing unparalleled research and innovation opportunities. However, developing such a platform would require significant leadership, resource and engagement with clinical, technical and research experts.

115 See https://shtg.scot/our-advice/a-national-radiology-information-system-ris-for-scotland-perceived-benefits-and-constraints-to-implementation/.

116 See https://dhcw.nhs.wales/product-directory/dataand-information/welsh-radiology-information-system-wris/.

117 See https://www.health-ni.gov.uk/news/advanced-digital-imaging-archive-improve-patient-outcomes.

118 See https://webarchive.nationalarchives.gov.uk/ukgwa/20210401201200/https://improvement.nhs.uk/documents/6119/Transforming_imaging_services.pdf.

In the meantime, outside Scotland, several individual hospitals or groups of hospitals have developed or are developing data resources, based on routine NHS radiology imaging activity, to support research and analysis for public benefit. Some of these have benefited from substantial government investments in imaging AI capability.[119] Multi-site NHS imaging databases, developed to support research for patient and public benefit, have generally focused on specific imaging modalities, sites and health conditions. A few illustrative examples are shown in Box 3.7. These resources show that it is possible to bring together imaging data from several hospitals. But they also highlight the many challenges of integrating and providing access to images using existing systems.[120] None of them has yet demonstrated a mechanism that could scale efficiently across the whole population of England for multiple imaging types in a way that could revolutionise understanding of the imaging characteristics and evolution of the full range of health conditions investigated with radiology imaging in the NHS.

119 E.g. see https://www.gov.uk/government/news/funding-boost-for-artificial-intelligence-in-nhs-to-speed-up-diagnosis-of-deadly-diseases.

120 See https://journals.sagepub.com/doi/pdf/10.1177/20552076211048654.

Box 3.7 Multi-site NHS imaging databases supporting research

- OPTIMAAM: A mammography imaging database to support the development and adoption of advanced tools (including AI) for early detection of breast cancers in the NHS Screening Programme. It includes over two million images from serial screening mammograms covering a 10-year period from over 170,000 women attending three centres.[121]
- ORFAN: A database of routine cardiac CT scans from over 40,000 people scanned at eight UK hospitals. It has been used to help develop and evaluate an AI tool that detects changes in arteries on the scans to predict the future risk of heart attacks.[122]
- NCCID: The National COVID-19 Chest Imaging Database of chest X-ray, CT and MRI images from over 20,000 patients with COVID-19 attending hospitals across the UK (mainly in England). Hosted by the NHS AI Lab, the database was rapidly established in 2020 through a partnership between the NHS, universities and industry. It aims to improve understanding of COVID-19 through studying how it affects chest images, and to develop ways of analysing images that improve care for patients hospitalised with COVID-19.[123]

Eye imaging: tackling complex challenges to develop large-scale retinal image resources

Clinicians and scientists increasingly recognise that features of the retina detected in retinal images[124] can provide early warning signs not only of eye diseases, but also of a range of other health conditions, including heart disease, stroke, dementia and Parkinson's disease. This means that large collections of retinal images acquired in NHS settings, linked to other sources of health data providing information on health outcomes, could be used for the development and testing of automated systems (including AI) to identify asymptomatic people at high risk of developing eye and other health conditions. This is of great interest because retinal imaging is fast and inexpensive (compared, say, with brain imaging) and non-invasive.

Retinal photographs from community settings

Thousands of retinal photographs are captured daily during NHS eye examinations by optometrists at high street opticians across the UK, where millions of eye tests are conducted each year.[125] Most of these photographs remain within the many proprietary systems used to capture and store them, representing an untapped resource for research and innovation. This major potential public health benefit is driving an ongoing initiative to create a national retinal imaging research resource in Scotland (Box 3.8).

121 See https://medphys.royalsurrey.nhs.uk/omidb/.

122 See https://oxfordbrc.nihr.ac.uk/ai-tool-could-help-thousands-avoid-fatal-heart-attacks/.

123 See https://transform.england.nhs.uk/covid-19-response/data-and-covid-19/national-covid-19-chest-imaging-database-nccid/ and https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8633457/pdf/giab076.pdf.

124 The two commonest types of retinal images are retinal photographs and optical coherence tomography scans. Retinal photographs are acquired in both community and hospital settings, while optical coherence tomography scans are only acquired in specialist hospital settings.

125 E.g. over 13 million per year in England and over 2 million per year in Scotland.

Box 3.8 Scotland's plans for a national retinal imaging resource to improve eye care and research

The Scottish Collaborative Optometry-Ophthalmology Network e-research (SCONe) aims to create a national retinal image resource in Scotland to enable early identification of eye disease, improve clinical outcomes and uncover novel ways to predict eye and other diseases. The ambition is to create copies of all the retinal photographs captured during routine eye tests in almost 1,000 community-based optometry practices across the country, and to link these with other national sources of health data.[126]

Given the hundreds of optometry outlets and many different computer systems involved,[127] scaling such an endeavour across the population poses substantial challenges, even for a relatively small country such as Scotland. This contrasts with the situation of working with a single Picture Archive and Communications Systems provider to enable creation of the Scottish Medical Imaging resource (Box 3.6). Optometry practices are not currently required to make available to the NHS the retinal images and associated clinical data they acquire through providing NHS eye services. Including such a requirement in future contractual arrangements would avoid the current need to seek agreement from each optometry practice to provide the data they hold, overcoming one part of the scalability challenge.

Retinal images from hospital settings

Eye examinations of a subset of patients requiring specialist ophthalmology assessment are carried out in hospital settings. The images are held in many different hospital-based systems, and there are no national, population-wide hospital NHS eye imaging resources or collections. However, there are some large, regionally based initiatives collating retinal images, as illustrated in Box 3.9.

Box 3.9 Curation of retinal images by the INSIGHT Health Data Research Hub

Health Data Research UK's INSIGHT health data hub has curated sets of retinal images from two large specialist eye centres in London and Birmingham.[128] The largest set includes over 13 million eye images from 320,000 patients who attended routine specialist outpatient appointments and ophthalmic accident and emergency at Moorfields Eye Hospital sites in London. Some of these images have been linked to national sources of health data and used to develop and test innovative AI approaches that could help to improve diagnosis of eye diseases, such as diabetic retinopathy and glaucoma, as well as to determine the risk of developing health conditions such as Parkinson's disease, stroke and heart failure.[129]

126 See https://www.ed.ac.uk/clinical-sciences/ophthalmology/scone/about-scone.

127 There are >7,000 optometry outlets providing NHS services across the UK (5,800 in England, 800 in Scotland, 350 in Wales, 270 in Northern Ireland). See: https://optical.org/media/hodlzrvn/ee-mapping-of-optical-businesses-final-report-22-feb-2023.pdf.

128 See https://www.insight.hdrhub.org/datasets.

129 See Zhou Y et al. *A foundation model for generalizable disease detection from retinal images.* Nature 2023. (https://www.nature.com/articles/s41586-023-06555-x).

Digital histopathology imaging: regional expertise with opportunities to scale nationally

When patients have a biopsy or a surgical procedure to remove abnormal tissue, such as a tumour, tissue samples are sent to specialist pathology laboratories. The samples are processed and prepared by technicians. Tiny slivers of tissue are mounted on glass slides. These are then either examined under the microscope by specialist histopathology doctors or – increasingly, as more modern systems are introduced – scanned into computer systems and examined and stored as digital images.

Rapid advances in digital pathology in the last 15–20 years have ushered in a digital revolution for histopathology. Digital image capture, storage, viewing, reporting and sharing has fast become the norm among digitally mature healthcare providers internationally. Investments in centres of pathology digital imaging excellence in the UK mean that specific regions of the UK are leading the way,[130] but the digitisation of pathology tissue imaging in the NHS is many years behind radiology and still far from complete nationally.

As with radiology, the sharing of images between NHS tissue pathology laboratories should enable expert clinical discussions and sharing of reporting tasks, which is of huge importance for a greatly overstretched speciality. Digitised pathology images are handled within PACS systems, and – as with radiology PACS – seamless interoperability of these systems is essential for streamlined sharing, reporting and analysis of images.[131]

As for radiology imaging, there is substantial potential for the development, evaluation and implementation of AI classification of images to enhance the efficiency and accuracy of specialist reporting, with increasing examples of automated image analysis being as good or even better than assessment by human experts. For example, AI models for digital pathology imaging have demonstrated excellent accuracy in identifying breast cancer metastases in lymph nodes or subtle characteristics indicative of skin cancers and in detecting melanoma or predicting prostate cancer progression.[132]

Many of the technical challenges (including storage, transfer, de-identification, image format standardisation, linkage, secure access and analysis) are similar to those for digital radiology. However, digital pathology images have much greater data volumes.[133] Addressing these challenges will require partnerships with companies developing PACS and other software systems that provide solutions. These will underpin the roll-out of digitisation of pathology across the NHS. They will also be needed to enable pathology imaging data access, linkage and analysis capabilities within NHS secure data environments to support the ongoing development and testing of new AI capabilities in a fast-evolving field.[134]

130 E.g. the Leeds-based National Pathology Imaging Collaborative (https://npic.ac.uk/about-us/), which aims to grow rapidly to include over 40 hospitals in the NHS, scanning over 3 million images per year, creating the biggest national digital pathology network in the world.; the Coventry/Warwick-based PathLAKE consortium (https://www.pathlake.org/).

131 Indeed, in 2022, Northern Ireland became the first UK region to combine pathology and radiology images and reports in the same PACS system (NIPACS – see https://www.healthtechdigital.com/northern-ireland-digitises-pathology-with-sectra/), embracing the opportunities for shared learning and digital infrastructure across radiology and pathology imaging NHS specialties.

132 See Kiran N et al. *Digital Pathology: Transforming Diagnosis in the Digital Age.* Cureus 2023 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10547926/).

133 E.g., NPIC (https://npic.ac.uk/about-us/) estimated that15 NHS sites serving 6 million people across Yorkshire and the North East of England would generate over 2.4 million pathology images (3 petabytes of image data) per year. This contrasts with 3 petabytes of NHS radiology imaging data from a 10-year period for the whole of Scotland, a similar-sized population (see earlier paragraphs on radiology imaging in this section).

134 See https://www.ukri.org/wp-content/uploads/2023/11/IUK-241123-DataEarlyDiagnosisPrecisionMedicineChallengeInteroperabilityRecommendations.pdf.

3.1.8 Screening data

Similar national screening programmes across the four nations

Under the guidance of the UK National Screening Committee, there are eleven national population screening programmes in England,[135] with similar (although not identical) screening programmes in Scotland, Wales and Northern Ireland.[136] These cover cancer screening (breast and cervical in women, bowel in men and women), screening for abdominal aortic aneurysms in men, screening during pregnancy in women for conditions affecting the unborn baby, and neonatal screening programmes.[137] Data on screening invitations, attendances and screening test results are collected and collated nationally in each country for each screening programme. The screening programmes and national screening data are managed in-house by national NHS organisations for some screening programmes and by commercial providers for others.[138] The screening programme data are used to monitor screening uptake, allowing modification of the programmes if needed.[139]

Improvements in screening impeded by difficulties linking screening to other national health data

Screening programme leaders, expert advisers, policymakers and researchers told us that, at least in England, it has been slow or sometimes impossible to link these screening programme data to other, existing national sources of health data. Like many other health data access and linkage challenges, this is mainly due to difficulties in navigating the legal and regulatory requirements for data sharing, compounded by limited capacity in the teams working to provide the data. Resolving these issues is important because linking to other health data is needed to find out whether the screening programmes are working as they should (i.e. detecting disease early and improving overall health outcomes), and to assess different screening approaches. For example, breast, cervical and bowel cancer screening programme data are not linked routinely to national cancer registry, cancer treatment, hospital, general practice and death registry data. If these data were linked and made securely available, analyses could be done to assess which people who have or have not been screened go on to develop cancer, and what their subsequent health outcomes are. Such linked data could also be used to investigate improved approaches to screening by targeting screening at those people most likely to benefit.

135 See https://www.gov.uk/government/collections/population-screening-programmes-document-collection.

136 See: https://www.gov.uk/guidance/screening-programmes-across-the-uk.

137 See: https://www.gov.uk/guidance/population-screening-explained.

138 See: https://digital.nhs.uk/services/screening-services and https://www.gov.uk/guidance/principles-of-population-screening/it-and-data.

139 E.g. see: https://www.england.nhs.uk/statistics/statistical-work-areas/screening/; https://www.scotpho.org.uk/health-conditions/screening/data/; https://phw.nhs.wales/services-and-teams/screening/; https://cancerscreening.hscni.net/.

3.1.9 Mental health data

Across all four nations of the UK, data from general practice records and national hospital inpatient and outpatient data (sections 3.1.2 and 3.1.4) include substantial quantities of data on mental health diagnoses, and – in the general practice records – on symptoms, signs, referrals and treatments, including medicines. Additional information on medicine use relevant to mental health conditions is captured through the various sources of medicines data already discussed (section 3.1.5).

NHS England collects two additional national datasets about mental healthcare and services from community and hospital provider organisations (for example NHS mental health, learning disabilities and care trusts, NHS acute hospital trusts, and some independent and voluntary sector providers) across England. The first of these, the Mental Health Services Data Set (MHSDS), includes structured patient-level data from 2019 onwards about children, young people and adults who are in contact with services for mental health and wellbeing, learning disability, autism or other neurodevelopmental conditions. The data include information on diagnoses, care, services and treatments. Some items overlap with data already collected through the general practice, hospital and medicines data sources discussed earlier (sections 3.1.2 and 3.1.5).[140]

The second, the Improving Access to Psychological Therapies dataset, includes data on referrals for and provision of NHS 'talking therapies' for anxiety and depression.[141] Like the MHSDS, it combines data items not already collated at national scale with items already available within other national data collections.

Both these datasets can be made securely available for a range of uses, including service planning, audit and research, and can be linked at individual patient level to other sources of health data.

140 See https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/mental-health-services-data-set and https://www.datadictionary.nhs.uk/data_sets/clinical_data_sets/mental_health_services_data_set.html#dataset_mental_health_services_data_set.specification.

141 See https://www.datadictionary.nhs.uk/data_sets/clinical_data_sets/improving_access_to_psychological_therapies_data_set.html.

3.1.10 Maternity and neonatal data

As for mental health data, a substantial amount of NHS data generated through the care of women through pregnancy and childbirth are available within existing nationally collated data sources, notably general practice records, hospital data (which include information on maternity and neonatal admissions) and medicines data. In England, an additional routine national collection of NHS data, the Maternity Services Data Set, was established by NHS Digital to supplement these. This is a structured, coded patient-level dataset for the whole of England that captures key information from each stage of the maternity care pathway including mothers' demographics, booking appointments, admissions and re-admissions, screening tests, labour and delivery. It also includes newborn babies' demographics, admissions, diagnoses and screening tests.[142]

The data are linkable at patient level to other health data sources. Importantly, they link data from mothers to their babies.[143] Similar maternity and newborn baby linked datasets are collected nationally in Wales (the Maternity Indicators Data Set since 2016)[144] and Northern Ireland (Northern Ireland Maternity System)[145] and are in advanced stages of development in Scotland.[146]

As regards digital maturity and interoperability in NHS neonatal and maternity care, there has been increasing uptake over the years of EPRs provided by the BadgerNet system across all four nations of the UK. Almost all neonatal care and an increasing proportion of maternity care in the UK is now recorded in BadgerNet.[147] This may help with system interoperability and data sharing across the UK for neonatal and maternity care but does not fix the substantial challenges of interoperability with hospital and general practice EPR systems more broadly. The provision of neonatal electronic patient records by a single system has also facilitated the data collection and curation efforts of the National Neonatal Research Database, which is covered further in section 3.1.12.

Great benefit can be gained from linking national maternity and neonatal data to a range of other health data. For example, this could support analyses to generate insights into the impact of neonatal characteristics and care on neonatal and later life health outcomes.

142 See https://digital.nhs.uk/data-and-information/areas-of-interest/maternity.

143 Linking data from mothers to data from their babies is crucial for being able to use data to understand the impact of the health and care of mothers before and during pregnancy on the health outcomes of babies and children. This link can be created by applying specially developed algorithms to hospital episodes data (see Harron K et al. Linking data for mothers and babies in de-identified electronic health data. *PLoS One* 2016 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0164667).

144 See https://www.gov.wales/maternity-and-birth-statistics-quality-report-html.

145 See https://healthdatagateway.org/en/dataset/19.

146 See https://perinatalnetwork.scot/data/matneo-data-hub-workstreams/.

147 See https://www.healthcareittoday.com/2023/02/24/clevermed-and-its-badgernet-solution-join-the-system-c-family/.

3.1.11 Patient-reported outcomes data

Since patients are experts in their own health, their own reports of their health through patient-reported outcome measures (PROMs) provide important information on their health status that can be used to evaluate and improve a wide range of health interventions and services. PROMS enable the measurement of many of the outcomes that matter most to patients, such as their quality of life. They are increasingly used to assess the effectiveness of interventions in clinical research studies and clinical trials. Their collection in such research studies may use innovative data collection methods, such as mobile phone apps (see section 3.5.1).

However, there are few national NHS PROMS data collections. For example, a search for 'patient-reported outcome measures' on the Health Data Research UK Innovation Gateway[148] in August 2024 identified only five datasets. Only three of these were national datasets, and all in fact different versions of the same NHS England PROMS dataset. NHS England's PROMs data have collected and regularly reported information on the health gain reported by patients undergoing hip or knee joint replacement procedures in the NHS in England since 2009.[149] The data can be linked at person-level to other datasets (for example hospital episode statistics) collected by NHS England.[150]

3.1.12 National audits and registries

Many national audits and registries have complex funding, commissioning, hosting and access arrangements

Many national datasets fall under the general label of 'audits and registries'. There are well over 100 of these but no unified catalogue. They comprise national collections of data that monitor the occurrence and management of specific health conditions (for example cancer, heart disease, stroke, diabetes and asthma), or specific procedures (for example knee and hip joint replacements or heart valve replacement procedures). Some span more than one of the four UK nations; others cover a single country only. Some have been established for decades; others have been set up much more recently. They are funded, established, managed, commissioned, hosted and delivered by a range of different organisations. These include national NHS organisations (such as NHS England[151] or Public Health Scotland[152]), health quality and improvement commissioning bodies (for example England's Health Quality Improvement Partnership[153]), national medical royal colleges,[154] independent data registry service providers,[155] large and smaller charity research funding organisations (for example

148 See https://www.healthdatagateway.org/.

149 See https://digital.nhs.uk/data-and-information/publications/statistical/patient-reported-outcome-measures-proms.

150 See https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/patient-reported-outcome-measures-proms.

151 See https://digital.nhs.uk/data-and-information/clinical-audits-and-registries and https://digital.nhs.uk/services/national-disease-registration-service.

152 See https://publichealthscotland.scot/services/scottish-national-audit-programme-snap/scottish-cardiac-audit-programme-scap/ and https://www.publichealthscotland.scot/our-areas-of-work/disease-registration-and-screening/disease-registration/.

153 See https://www.hqip.org.uk/.

154 E.g. see https://www.rcseng.ac.uk/standards-and-research/support-for-surgeons-and-services/audit/national-audit/.

155 E.g., NEC Software Solutions UK (formerly Northgate Solutions, https://www.necsws.com/registries/) supports the National Joint Registry and the National Vascular Registry, both of which are commissioned by HQIP on behalf of NHS England.

British Heart Foundation,[156] Cystic Fibrosis Trust[157]), independent registered charities established specifically to host and run national audit programmes (for example Intensive Care National Audit and Research Centre[158]), health professional bodies (for example UK Renal Registry hosted by UK Kidney Association[159]), universities and NHS trusts (for example National Neonatal Research Database, hosted by the Neonatal Data Analysis Unit at the Chelsea and Westminster Hospital Campus of Imperial College London[160]). There are many different arrangements for commissioning, funding and delivery, with often complex, confusing and tortuous governance of data access. A recent summary from NHS England lists many – but not all – of the national disease audits and registries relevant to England.[161]

Many national audits and registries are managed or commissioned by national NHS organisations. Some examples of different NHS England-managed audits illustrating different arrangements for data collection, management, governance and access are shown in Box 3.10.

156 E.g., the NHS England Out of Hospital Cardiac Arrest Outcomes Audit (https://warwick.ac.uk/fac/sci/med/research/ctu/trials/ohcao/) is funded by the BHF and Resuscitation Council UK, and hosted by the University of Warwick and National Ambulance Service; the NHS England National Audit of Cardiac Rehabilitation (https://www.cardiacrehabilitation.org.uk/site/about-us.htm) is funded by the BHF (https://www.bhf.org.uk/informationsupport/publications/statistics/national-audit-of-cardiac-rehabilitation-quality-and-outcomes-report-2021), commissioned through NHS Arden & Gem, and hosted by the University of York with informatics and data management services provided by NHS England (https://digital.nhs.uk/data-and-information/clinical-audits-and-registries/national-audit-of-cardiac-rehabilitation).

157 The UK Cystic Fibrosis Registry is hosted and managed by the Cystic Fibrosis Trust – see https://www.cysticfibrosis.org.uk/about-us/uk-cf-registry.

158 See https://www.icnarc.org/.

159 See https://ukkidney.org/about-us/who-we-are/uk-renal-registry.

160 See https://www.imperial.ac.uk/neonatal-data-analysis-unit/neonatal-data-analysis-unit/contributing-to-the-nnrd/.

161 See https://www.england.nhs.uk/publication/list-of-national-clinical-databases-registries-and-audits/.

Box 3.10 Examples of NHS England audits with different arrangements for data collection, management, governance and access

CVD Prevent[162]

This is a national audit of cardiovascular prevention in primary care. It is funded by NHS England and commissioned by the Healthcare Quality Improvement Partnership (HQIP). It is based on analyses of routinely collected general practice data, extracted via the NHS England General Practice Extraction Service (GPES). There is currently no mechanism for potential users not directly involved in the audit to access these data for broader beneficial uses. However, many of the data items are available for COVID-related research as they are included within the GPES General Practice Data for Pandemic Planning and Research.

The National Disease Registration Service

This includes NHS England's National Cancer Registration and Analysis Service (NCRAS) and the National Congenital Anomaly and Rare Disease Registration Service (NCARDRS) population-based registers.[163] The long-established national cancer registration system has benefited in previous years by moving from a multi-regional to a consolidated national approach. It has also been enhanced with increasingly rich data on cancer diagnoses, including stage, grade, genomic information, treatment information (chemotherapy and radiotherapy) and cancer waiting times. However, data availability has always lagged behind real time by up to a year or more.

Further, access to these data, previously via Public Health England, was negatively affected by loss of expert staff and the difficulties of integrating different data platforms following the dissolution of Public Health England in September 2021 and NHS Digital's merger with NHS England in February 2023. Recovery from these negative impacts is underway but not yet complete.

The National Diabetes Audit

This audit of the care of patients with diabetes is commissioned and managed by HQIP on behalf of NHS England and the Welsh Government.[164] In England, core data items are extracted from general practice data systems via NHS England's General Practice Extraction Service. Further data items are incorporated from secondary care and diabetic retinopathy screening services. Although listed as a dataset that can be requested via NHS England's Data Access Request Service,[165] in practice access to data for beneficial uses by researchers and analysts not directly involved in the audit programme has, for many years, proved to be very challenging – and in many cases impossible.

162 See https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof/cardiovascular-disease-prevention-audit-cvdprevent and https://www.cvdprevent.nhs.uk/.

163 See https://digital.nhs.uk/ndrs/ and https://digital.nhs.uk/ndrs/our-work/ncras-work-programme.

164 See https://digital.nhs.uk/data-and-information/clinical-audits-and-registries/national-diabetes-audit/core.

165 See https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services/data-set-catalogue/national-diabetes-audit-nda.

NICOR cardiovascular audit and registry datasets

Most of these secondary care cardiovascular audits are commissioned by HQIP on behalf of NHS England. Several also include Wales and Northern Ireland. The audits are delivered by the National Institute for Cardiac Outcomes Research (NICOR).[166] They include audit data on heart attacks, heart failure, adult cardiac surgery, coronary interventions, congenital heart disease, cardiac rhythm management and heart valve replacement or repair procedures. Access to data from these rich and diverse registries, linked at national scale to other health data sources, for beneficial purposes beyond the audits themselves, has been very difficult for many years. Reasons have included limited data curation resources and complex data access and governance processes. Incorporation of several NICOR datasets into NHS England's SDE in partnership with Health Data Research UK's BHF Data Science Centre[167] has underpinned substantial progress in curation, linkage and access. Ongoing commitment from NHS England, the Department of Health and Social Care, HQIP, NICOR, Health Data Research UK and others will be needed to ensure these advances are sustained beyond the pandemic.

These few examples demonstrate the pressing need for rationalisation and consolidation of the UK's national disease audit and registry datasets. While the close involvement of expert, specialist groups in the collection and curation of these datasets is crucial, far more streamlined and cost-effective systems are urgently needed for their collection, access, linkage and analysis for broader benefit. In England, this could potentially be provided via the NHS England Outcomes and Registries Programme.[168] This focuses on implantable device registries in order to address the need for a more consistent approach to assure safety and satisfactory outcomes, highlighted in Baroness Cumberlege's Independent Medicine and Medical Device Safety Review in 2020.[169] However, the remit, resourcing and expertise of the programme would need to be extended well beyond its current focus on implantable device registries for this to be a workable solution.

166 See https://www.nicor.org.uk/.

167 See https://bhfdatasciencecentre.org/areas/cvd-covid-uk-covid-impact/.

168 See https://www.england.nhs.uk/outcomes-and-registries-programme/.

169 See https://immdsreview.org.uk/Report.html.

3.1.13 Operational and workforce data

Operational data

Some operational NHS datasets, for example on ambulance call out times, emergency department waiting times, hospital bed occupancy and numbers of people on waiting lists for treatments, are collected and collated nationally.[170] Individual general practices and hospital trusts have access to much more detailed, real-time information, for example about the sources and reasons for delayed discharge of hospital inpatients or operating theatre capacity. Collecting accurate, high-quality operational data, both regionally and nationally, and using it to inform service improvement, is critical to the smooth running of the health service. Individual datasets are routinely used in this way.[171]

Going beyond individual dataset analyses by linking these data to other sources of health data, for example on health outcomes, would allow the evaluation of the impact of operational factors on service users' health, including revealing and mitigating against any inequalities. Such linkages are not routine. However, secure, timely access to relevant linked data would allow researchers with data analysis skills and expertise in health services, health economic and policy research to better inform healthcare policy. For example, they could independently evaluate the costs, cost effectiveness and health benefits of new AI data-driven methods that aim to optimise hospital bed usage, reducing delays to discharges and the length of hospital stays.

Workforce data

National NHS bodies collect person-level data about all NHS employees, including their basic characteristics (age and sex), job type, workplace (which hospital, general practice or other healthcare setting), working time commitment and salary. The primary purpose of collecting these data is to monitor, compare and understand costs nationally and regionally, seeking improvements and potential efficiency gains.[172] This is because pay for the NHS's large number of staff (for example 1.4 million NHS staff in England) is the single largest cost within the healthcare system.

Because these data are collected at the individual person level it is technically possible to link them to other sources of health data. This would allow analyses of health risks and outcomes among healthcare workers. However, such linkages are not routinely or readily conducted across all nations of the UK. During the pandemic, health and care workers were among those across society exposed to some of the highest risks to their personal health.[173] At the start of the pandemic, Scotland and Wales already had secure systems for linkage of and access to these types of data. This made it possible to link records between workforce, COVID-19 testing and health outcomes data early during the pandemic in Scotland to evaluate the risk of severe COVID-19 among healthcare workers in patient- and non-patient facing roles, and their household members.[174] This type of analysis should not be confined to the UK's devolved administrations or to COVID-19-related

170 E.g. see https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets.

171 E.g. see https://www.england.nhs.uk/long-read/operational-performance-update-oct-23/.

172 E.g. see https://digital.nhs.uk/data-and-information/areas-of-interest/workforce.

173 E.g. see ONS report *COVID-19 related deaths by occupation in England and Wales,* 2021: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/bulletins/coronaviruscovid19relateddeathsbyoccupationenglandandwales/deathsregisteredbetween9marchand28december2020 and BMA report *The impact of the pandemic on the medical profession,* 2022 https://www.bma.org.uk/media/2jhfvpgk/bma-covid-review-report-2-september-2024.pdf.

174 See Shah ASV et al. *Risk of hospital admission with coronavirus disease 2019 in healthcare workers and their households: nationwide linkage cohort study.* BMJ 2020 (https://www.bmj.com/content/371/bmj.m3582).

investigations. In today's NHS, where recruitment and retention of our healthcare workforce is a major challenge, it is crucial that workforce data linked to other health data sources can be used to support a wide range of studies of health outcomes among healthcare workers in different roles. Insights from these could then inform policies to reduce work-related physical and mental health problems, and so support a thriving and healthy healthcare workforce.

3.1.14 Data from private healthcare providers

Most healthcare (82%) in the UK continues to be funded by the NHS.[175] This is especially so for emergency and unscheduled care, where private provision remains extremely rare. The NHS purchases some healthcare from commercial providers. This includes many community-based eye tests, dental assessments and treatments, some radiology imaging tests, and certain scheduled procedures and operations such as cataract extractions and joint replacements. Further, some assessments, tests and procedures are both funded and provided privately, outside the NHS, through medical insurance schemes or direct out-of-pocket payments. The proportion of healthcare funded in this way has increased in recent years, partly due to restrictions in eligibility for NHS funding of certain types of care (for example eye and dental assessments in the community) as well as excessively long waiting lists for specialist assessments, investigations and procedures, particularly following the COVID-19 pandemic. (Also see sections 1.1 and 3.1.3 and their relevant footnotes).

Where providing data to national data collections is needed for the processing of payments, financial incentives mean that returns are generally near 100% complete. For example, this applies to data collected by the NHS Business Services Authority on NHS-funded community eye tests and dental assessments as well as on community dispensed medicines. As discussed earlier on retinal imaging (see section 3.1.7), the provision of additional, richer data generated as part of the healthcare service (such as retinal images) could be included within contractual requirements. This would substantially enhance existing national data collections and their use at scale in a wide range of health service planning and research analyses.

Data from non-NHS-funded private sector health services (funded through insurance schemes or direct payment) are not currently included in national and regional data resources and collections. Information about some of this private healthcare may find its way into NHS general practice data records, but this is unlikely to be complete. If non-NHS-funded private healthcare provision continues to grow, in line with the current direction of travel, the 'data gap' looks set to widen, particularly for scheduled tests and procedures that are attractive business opportunities for private providers. However, the relevant data exist within the EPRs and other computer systems of the relevant private healthcare companies. In England, the Acute Data Alignment Programme aims to adopt common standards for data collection and performance measures across the NHS and private healthcare, and to support work towards the direct submission of data from private providers to NHS England.[176] One way to ensure that data from privately funded healthcare are provided to national NHS data systems would be through legislation. For example, mandatory provision of data supports the assembly of national population-based linked datasets across Australia.[177]

175 See https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthcaresystem/bulletins/ukhealthaccounts/2022and2023.

176 See https://digital.nhs.uk/services/acute-data-alignment-programme.

177 See Smith M and Flack F. *Data Linkage in Australia: The First 50 Years.* Int J Environ Res Public Health 2021 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8583508/).

3.2 Health-relevant administrative data arising outside the healthcare system

3.2.1 Birth and death register data

All births and deaths in the UK must be registered by law, within 42 and five days respectively in England, Wales and Northern Ireland, and within 21 and eight days respectively in Scotland. The data from birth and death certificates are recorded electronically in national birth and death registries, held and managed separately by the General Register Offices for England and Wales, the General Register Office for Northern Ireland, and the National Records of Scotland for Scotland.[178] Birth registers include information on the name, date and place of birth of each individual, the name and place of birth of the mother, and the name and place of birth of the father, if he is included on the birth certificate. Death registers include information on the name, date and place of death of each person who has died, their age at death, last occupation, final residence and cause of death.

National birth and death registration information is shared with the Office for National Statistics (ONS) (or equivalent national statistics bodies in Scotland and Northern Ireland) by the General Register Office for England and Wales (or equivalent organisations in Scotland and Northern Ireland). This enables the production of national birth and mortality statistics. These in turn inform service planning and resource allocation, estimates and projections of population numbers and life expectancy, analyses of socio-demographic trends, patterns and trends in specific causes of death, and a range of other analyses in the public interest.[179] Birth and death data can be linked to other health-relevant data and can be used in a wide range of health research. For example, such linked data can be used to discover the dates and causes of death of participants in a clinical trial assessing the benefits and harms of a new treatment, or to study the relationship between prior health conditions, healthcare provision and the timing and causes of death in the population. National death registry data are provided regularly by the ONS (or equivalent national statistics bodies) to national health data custodians (for example NHS England) in each of the four nations of the UK to support healthcare provision, planning and research.

Deaths that occur in hospital are also recorded in national hospital episodes statistics (but these do not include deaths occurring outside hospital), while general practices also record information in their electronic health records on the deaths of patients in each practice. National Health Service midwives use the birth notification process to record information on all births that the NHS knows about (the vast majority), and to ensure that all newborns are assigned an NHS number (or community health index (CHI) number in Scotland).[180] While not identical, NHS birth notification data and birth registration data are very similar.

While most deaths are registered promptly (for example, in England and Wales over 93 % of deaths were registered in the same year as they occurred), there are concerns about the rising proportion of late registrations in England, Wales and Northern Ireland, where registration of deaths that are subject to an inquest are delayed until the cause of death has been established. Delays for inquest verdicts on suicides and drug-related deaths cause inaccuracies in the estimation of calendar-year trends.[181]

178 See https://www.gov.uk/general-register-office.

179 See https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/sourcesofdata/generalregisterofficegro.

180 See https://digital.nhs.uk/services/national-care-records-service/birth-notification-process.

181 See https://rss.org.uk/news-publication/news-publications/2019/general-news-(1)/rss-highlights-late-death-registrations-problem-to/.

3.2.2 Social care data

Need for digitisation across the social care sector

As discussed earlier, the social care sector lags behind healthcare in terms of digital maturity (see Chapter 1). Introducing digital systems into social care will be an essential step towards data integration across social care and with healthcare providers. Such integration has been a stated ambition of governments across the four UK nations for many years. This vision needs to be matched by appropriate investment and delivery. For example, the implementation of digital social care record systems in England has been accelerated by investment through the Adult Social Care Digital Transformation Fund.[182] Unlike the rest of the UK, Northern Ireland has for many years had combined health and social care trusts, but still needs improved digital maturity, especially in social care.

COVID-19 pandemic highlighted deficits in national social care data

Digitisation should also increase the efficiency and reduce the costs for social care provider organisations of providing data to mandatory national collections. Until recently, these have been very limited across all four nations of the UK. The COVID-19 pandemic highlighted that limitations in social care data have hindered service development and research for years, especially with respect to care homes.[183] This realisation led to calls for a frequent, person-level, national care home data collection, including all existing care home residents, irrespective of the source of their care funding. Such data collection would enable tracking of those admitted into or discharged from care home settings, linking via the Unique Property Reference Number[184] to identify care home locations as shared residences.

182 See https://transform.england.nhs.uk/key-tools-and-info/adult-social-care-digital-transformation/digitising-social-care-fund/.

183 See Burton JK et al. *Closing the UK care home data gap – methodological challenges and solutions.* International Journal of Population Health and Data Science 2020 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8138869/).

184 UPRN: a unique numeric identifier for every address across the UK (see https://www.geoplace.co.uk/addresses-streets/location-data/the-uprn).

Progress since the pandemic but improvements still needed

There has been some progress since the pandemic. In England, up to 2023, social care providers submitted data on adult social care in aggregate and only once per year. In 2023, a new system was introduced for collecting person-level data quarterly on local authority adult social care (Adult Client Level Social Care Data).[185] This aims to enable more timely and flexible analysis of adult social care provision at national and regional levels and – in due course – linkage to other national health datasets. However, the new data does not include information on self-funded, independently arranged adult social care.[186]

Linking these new adult social care data to other health data would be of great value, enabling people's care to be tracked across the health and social care system, at least where their care is NHS or local authority funded. Analyses of such linked data could generate insights on delays in care pathways and their costs, inform service planning, assess inequalities in the provision of care, and examine how different types of care affect health outcomes. Including these social care data among NHS England's Data Access Request Service datasets[187] would enable access and linkage to other data, avoiding the unnecessary costs and complexity of establishing a separate data access route.

In Scotland, Public Health Scotland collects national person-level data quarterly on adult social care clients and the services they receive. These data are linkable to other national sources of health data, which should enable the same types of analyses as outlined for English data.[188] In Wales, local authorities provide person-level data to the Welsh government about adults receiving social care in an annual census.[189] Some adult social care data in Northern Ireland are collected annually, currently in aggregate form.[190]

185 See https://www.ardengemcsu.nhs.uk/adult-social-care-client-level-data/.

186 See https://www.gov.uk/government/publications/adult-social-care-in-england-statistics-background-quality-and-methodology/adult-social-care-in-england-statistics-background-quality-and-methodology#annex-b-official-statistics-in-development---client-level-data.

187 See https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services/.

188 See https://publichealthscotland.scot/services/national-data-catalogue/national-datasets/search-the-datasets/social-care-source/.

189 See https://www.gov.wales/data-collection-local-authority-social-services.

190 See https://www.health-ni.gov.uk/topics/dhssps-statistics-and-research-social-services/social-care-statistics.

Different processes for children's social care data

In England, person-level data on children referred to or receiving social care and looked after children are submitted annually by local authorities to the Department of Education.[191] In Scotland and Wales, similar person-level data are submitted annually by local authorities to the Scottish and Welsh Governments respectively.[192] In Northern Ireland, aggregate data on children referred to or receiving social care and on looked after children are collected from each of the five health and social care trusts and reported on annually (or for some data quarterly or monthly) by Northern Ireland's Department of Health. In addition, person-level data are collected and reported on annually for looked after children and care leavers.[193]

In England, the difference in controllership of national adult and children's social care data may impact the ease of data sharing and linkage. Adult social care data are collected and controlled by NHS England. Linking them to health data also collected by NHS England should be relatively straightforward. Organisational boundaries make it less straightforward to link data held by NHS England (such as adult health and social care data) to data from other sectors. By contrast, children's social care data are collected by the Department of Education (as opposed to NHS England). This makes their linkage within the Office for National Statistics (ONS) to other data (such as on educational attainment) from the Department of Education as well as data from other government departments (for example youth offending data from the Ministry of Justice)[194] easier than their linkage to NHS data about children's health.

3.2.3 Administrative data from other government sources

Multiple sources of health-relevant data

The ONS collects, stores, processes and uses data from a wide range of sources. These data are used to produce statistics to guide local and national government policy, to inform citizens, and to support research for wider public benefit.[195] Most of these data come from sources outside the health and care system but many are relevant to our health. Although the ONS is a UK body, the geographic coverage of the data it collects depends on the source, and on the purpose and legal basis for collecting it. ONS data always include England, but do not always include all the devolved administrations. Key sources of ONS data are summarised in Box 3.11.

191 See https://www.gov.uk/guidance/children-in-need-census and https://www.gov.uk/government/publications/children-looked-after-return-2023-to-2024-guide.

192 See https://www.gov.scot/publications/about-childrens-social-work-statistics/ and https://www.gov.wales/data-collection-local-authority-social-services.

193 See https://www.health-ni.gov.uk/topics/dhssps-statistics-and-research/childrens-services-statistics.

194 See https://www.adruk.org/our-mission/our-impact/analysis-of-childrens-educational-childrens-social-care-and-offending-characteristics/.

195 See https://www.ons.gov.uk/aboutus/usingpublicdatatoproducestatistics/answeringyourquestionsaboutdata.

Box 3.11 Sources of Office for National Statistics data informing policy and research

Surveys

These are often conducted in collaboration with other national bodies, universities or specialist research organisations. For example, during the COVID-19 pandemic, the ONS COVID-19 infection survey gathered, analysed and presented data from regular coronavirus testing of over half a million people in private residential households across England, Wales, Northern Ireland and Scotland. The survey was delivered in partnership with the universities of Oxford and Manchester, the UK Health Security Agency (UKHSA), Wellcome and multiple partner laboratories.[196] The results provided robust information for policymakers on estimated numbers and percentages of people testing positive for COVID-19 across all four countries of the UK, including trends over time and geographic variation.

Census

The ONS runs the England and Wales census every 10 years, providing accurate estimates of all the people and households in England and Wales, together with information about these people to guide local and national government policy on public services. For example, to plan and ensure equitable access to health services, policymakers need to know about the age, sex, ethnic and socio-economic make up of society across the different geographic regions of the UK.

Administrative data from government departments and public bodies

The ONS receives and uses administrative data from various government departments and other public bodies, including the Department for Work and Pensions (DWP), the Home Office, HM Revenue and Customs (HMRC), the Ministry of Justice, local authorities, the Higher Education Statistics Authority, the General Register Office and NHS England. These organisations collect and provide information generated by people's interactions with public services such as the education, justice, benefits, tax and health systems. Collecting, linking and analysing this information can guide policy through providing policymakers with a detailed understanding of how people's health is influenced by and can impact educational opportunity and attainment, criminal behaviour and imprisonment, and financial status.

196 See https://www.gov.uk/government/news/covid-19-infection-survey-participants-thanked-for-huge-contribution-to-pandemic-response.

Many of these data are made available for approved uses by accredited external users via the ONS Secure Research Service (SRS).[197] This has been providing secure access to de-identified data for accredited researchers for over 15 years and is one of the largest secure data environments in the UK. The SRS is now gradually being replaced by a new service, the ONS Integrated Data Service, established to take advantage of rapid advances in technology to provide data and analytical and visualisation tools in a secure multi-cloud infrastructure.[198]

Data from similarly diverse sources are collected by national statistics organisations in the devolved administrations of Scotland, Wales and Northern Ireland. Partnership with and investment from Administrative Data Research UK has been critical to the growth in availability and wider uses of multiple sources of administrative data in all four nations of the UK, informing an increasingly wide range of research and analysis in the public interest.[199]

Linking health and care data to other sources remains difficult

Collecting, curating and enabling linkage of and access to multiple sources of data so that analyses can be conducted by approved and accredited researchers is a major undertaking. Linking and analysing data from the health and care system with data from other administrative sources is of major importance for understanding the wider determinants and consequences of good and poor health. For example:

- linking data on educational attendance and attainment to NHS data on health service use and health outcomes allows studies on how health can affect access to education and on how educational opportunities and attainment can affect later life health and wellbeing;
- linking data on employment status and earnings to health and disability data enables studies to better understand the health-related causes and consequences of the rising proportion of economically inactive people in the UK (22% of 16–64-year-olds in early 2024);
- linking data on criminal offenders and prisoners with data on benefit claims as well as data on health service utilisation and health conditions can support analyses to understand the relationship between imprisonment, reoffending and subsequent health and work status.

As a result of legal and non-legal hurdles, linking healthcare to non-healthcare data in the UK has continued to be slow, difficult or impossible. The exception is in Wales, where the SAIL Databank has facilitated these types of linkages and enabled access to researchers for some years.[200] Progress is being made in Scotland and Northern Ireland under the auspices of Research Data Scotland[201] and the Digital Health and Care Northern Ireland Data Institute,[202] but these initiatives do not yet support cross-sectoral studies as a matter of routine. And there are considerable ongoing challenges in England.

197 See https://api-ons.metadata.works/branding/assets/about.html.

198 See https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice/integrateddataservice/transitiontotheintegrateddataservice and https://integrateddataservice.gov.uk/.

199 See https://www.adruk.org/data-access/data-catalogue/.

200 E.g. see John A et al. *Association of school absence and exclusion with recorded neurodevelopmental disorders, mental disorders or self-harm: a nationwide, retrospective, electronic cohort study of children and young people in Wales, UK*. Lancet 2022 (https://pubmed.ncbi.nlm.nih.gov/34826393/).

201 See https://www.researchdata.scot/.

202 See https://dhcni.hscni.net/digital-strategy/data/.

Legal complexity is one of the reasons why enabling access to linked data for these types of studies is not straightforward. The key challenge is typically establishing that data can be used or disclosed in accordance with the common law duty of confidentiality (rather than any barriers imposed by data protection legislation). There are various routes to complying with the common law duty of confidentiality. These include:

- Obtaining patient consent.
- The Digital Economy Act 2017 (DEA). This has facilitated the sharing and linking of de-identified data by public authorities for accredited research purposes through its provision that disclosure under the DEA does not breach the common law duty of confidentiality. However, the DEA currently excludes sharing of data relating to the provision of health services or adult social care by NHS and other bodies with functions related to health and social care.[203]
- Section 251 of the National Health Service Act 2006 and its current Regulations, the Health Service (Control of Patient Information) Regulations 2002 ('COPI Regulations'). These allow the Secretary of State for Health to set aside the common law duty of confidence for defined medical purposes. These powers were effectively used during the COVID-19 pandemic through the issuing of 'COPI notices', which mandated data sharing for pandemic planning and research. Section 251 and the COPI notices also establish a process for the Health Research Authority (HRA) to approve the processing of confidential information for medical research purposes.

Legislative changes could simplify the legal route to cross-sectoral linkages. Including health and care bodies in the provisions of the DEA may be one way to reduce the barriers to some types of linkage and analysis of cross-sectoral health-relevant data for public benefit. However, there are known public sensitivities with including health data in the scope of the DEA and any amendments to the DEA will require careful consideration with meaningful input from patient and public representatives. Changes to the COPI Regulations could simplify and streamline the processes for enabling health and care organisations to use and share patient information for health and wider research purposes. Wider use of COPI notices mandating data sharing should be considered as a mechanism, with appropriate public engagement and transparency.

3.3 Data collected specifically for health research studies

3.3.1 Main types of clinical and population health research studies

Broadly speaking, there are two main types of clinical or population health research studies that recruit and study human participants and are relevant to this review.

Observational studies

Observational studies recruit and study participants who may be drawn from the general population or recruited because they have – or are at high risk of – a particular health condition. Many different scientific approaches and study designs are used, but among the most common and relevant here are **prospective longitudinal cohorts.** These recruit and follow people over years or decades, aiming to understand the causes and consequences of different health conditions and to develop new ways to predict, prevent, diagnose and treat them. Across the UK we estimate that there are hundreds or possibly thousands of such

203 See https://www.adruk.org/fileadmin/uploads/adruk/Documents/The_legal_framework_for_accessing_data_April_2023.pdf.

studies. They have huge variability in research focus, study size, geographic distribution, range of participant ages, ethnicities and socio-economic backgrounds, and the diversity and depth of data types collected (for example some include rich imaging data, physiological measures, or in-depth cognitive assessment). Increasingly, longitudinal cohorts collect bio-samples for laboratory analysis (for example blood, urine, saliva or tissue samples) from all or a subset of their participants, to add insights on biological mechanisms of diseases and their consequences. These studies also vary in the extent to which they are established as resources for access and use by a wide community of researchers, as well as in the extent to which they are embedded within the NHS. Several illustrative examples are shown in Appendix 6.

Intervention studies (trials)

These test the effectiveness and safety of a treatment or intervention, usually by comparing groups with and without the treatment or intervention. Most – although not all – well-designed intervention studies, are randomised clinical trials. These generate the gold standard evidence needed to gain regulatory approval for new therapies and to influence clinical or public health practice. Clinical trials must be registered in a public database,[204] such as clinicaltrials.gov,[205] searches of which show that there are thousands of UK-based clinical trials. Well known contemporary examples include the RECOVERY trial, testing treatments for people hospitalised with COVID-19 and other causes of pneumonia among 50,000 people, and the NHS-Galleri trial of a new blood test for early detection of cancer among 140,000 people.

3.3.2 Linking research studies to health and administrative records

Many contemporary health-related prospective cohort studies and clinical trials (including all the examples mentioned in section 3.3.1) aim to better characterise and follow the health of their participants through linkage to NHS and other health-relevant administrative data sources, usually with explicit participant consent. Such data linkage has significant advantages, both scientifically and for efficiency and cost-effectiveness. From a scientific standpoint, follow-up that requires the active engagement of participants, whether through face-to-face visits or through telephone, online or app-based questionnaires, is always affected by some attrition (or loss to follow-up), which can bias research results. Follow-up via linkage to health-related records places no further active engagement burden on the participants. As a result, follow-up through this route will be 100% complete, provided the linked records cover the relevant data sources and geographical distribution of the participants. Follow-up through linkage to routinely collected health and administrative records is also a potentially highly efficient and cost-effective way to follow the health of research participants, provided the relevant data exist, can be obtained, and are of sufficient quality and accuracy to be useful for the aims of the research.

Unfortunately, these important provisos are not always fulfilled. There are certainly plenty of data that could be useful for follow-up in many prospective longitudinal cohorts and clinical trials (see sections 3.1 and 3.2). Further, the quality and accuracy of several datasets have been shown to be sufficient for many research questions as well as demonstrating the necessary data provenance and integrity

204 See https://www.hra.nhs.uk/planning-and-improving-research/research-planning/research-registration-research-project-identifiers/.

205 See https://clinicaltrials.gov/.

for clinical trial regulatory purposes.[206] However, obtaining the data is often difficult. Long, frustrating and costly delays in the processes of requesting and accessing linked health systems data – or even failure to obtain linked data at all – are unfortunately all too common (see section 6.1 and Box 6.1).

3.3.3 Issues around consent

In almost all research studies of the types discussed in section 3.3.1, research participants give their consent to take part.[207] Obtaining explicit consent enables compliance with the common law duty of confidentiality. Consent is rarely – if ever – the lawful basis for access to health data for research under data protection laws.

Participants may commit significant amounts of time to completing questionnaires, having measurements and scans done, providing samples and attending various follow-up assessments. To protect the interests of potential research participants, before any proposed health and care research study can start, an appropriate research ethics committee must review and approve the overall aims, detailed plans, participant information and consent materials.[208]

Informed consent for research is widely discussed and debated by experts in philosophy, ethics, the law and research, as well as by patients and members of the public. Here, we touch briefly on those issues around consent that are most relevant to the linkage and use of routinely collected health, care and other relevant administrative sources in clinical and population health research studies.

When people consent to take part in a research study, it is important that they (or where relevant their proxy) understand what they are consenting to, the potential benefits (to them personally as well as more widely, for example for future patients) and any risks. For studies that require long-term follow-up over months, years, or decades, participants may – and could more commonly – be invited to provide consent for linkage to their health and administrative records.[209] They may also consent to the ongoing linkage and research uses of their data continuing after their death or in the event of future mental incapacity (for example, if they develop dementia). Consent for long-term follow-up is an important feature of prospective longitudinal cohort studies and some clinical trials. It is generally only after many years of follow-up, and after considerable investment of participant and study staff time and funding resources, that some of the most important research questions from longitudinal studies can start to be addressed. The benefits of long-term follow-up are well illustrated by studies of potential causes of dementia in the Million Women Study (Box 3.12)

206 E.g. see Sydes MR et al. *Getting our ducks in a row: The need for data utility comparisons of healthcare systems data for clinical trials.* Contemporary Clinical Trials 2024 (https://pubmed.ncbi.nlm.nih.gov/38537901/); Murray ML et al. *Data provenance and integrity of health-care systems data for clinical trials.* Lancet Digital Health 2022 (https://pubmed.ncbi.nlm.nih.gov/35868811/); Murray ML et al. *Demonstrating the data integrity of routinely collected healthcare systems data for clinical trials (DEDICaTe): a proof-of-concept study.* Health Informatics Journal 2024 (https://journals.sagepub.com/doi/full/10.1177/14604582241276969); Wilkinson T et al. *Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data.* Eur J Epidemiol 2019 (https://pubmed.ncbi.nlm.nih.gov/30806901/); Rannikmäe K et al. *Accuracy of identifying incident stroke cases from linked health care data in UK Biobank*. Neurology 2020 (https://pubmed.ncbi.nlm.nih.gov/32616677/).

207 A few exceptions are where parents provide permission on behalf of young children, where relatives or welfare guardians provide assent on behalf of adults who cannot do so for themselves (e.g. due to unconsciousness, dementia or severe mental illness), or where a healthcare professional decides that it is in the best interests of an incapacitated patient to take part in a clinical study, using an ethically approved waiver of consent procedure.

208 See Integrated Research Application System (IRAS): https://www.myresearchproject.org.uk/SignIn.aspx.

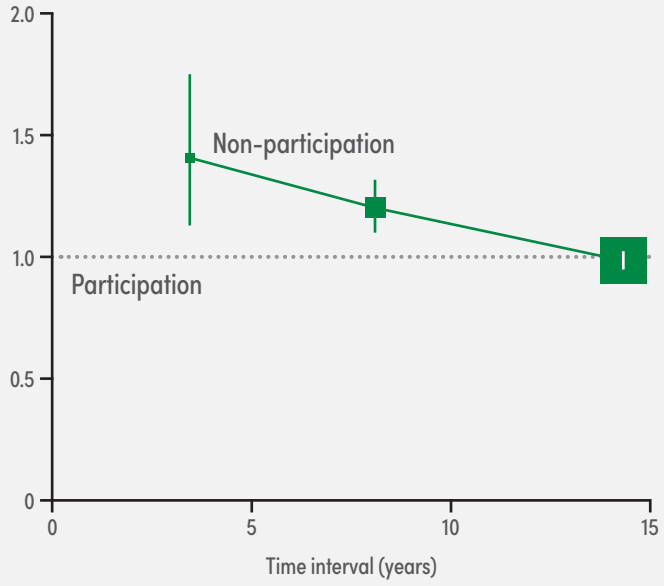
209 E.g., the internationally leading research resource, UK Biobank, recruited over half a million participants between 2006-2010 from across England, Scotland and Wales and has been following their health since (see www.ukbiobank.ac.uk).

Box 3.12 Insights into dementia from long-term follow-up in the Million Women Study

The UK Million Women Study is a population-based cohort of over a million women, recruited around 1998 and followed up by linking to their national NHS hospital records. In around 2001, most of these women answered questions about whether they took part in various cognitive and social activities: adult education; art, craft or music groups; and voluntary work. An average of 16 years later, hospital records showed that over 30,000 women had developed dementia. The risk of developing dementia was higher in women who had not participated in the activities, but only in the first 10 years of follow-up, and not thereafter.

This study demonstrates that long-term follow-up of a very large number of research participants can be achieved through linking to national routinely collected data. It would be impossible to follow the health of so many women for such a long time using any other method. It also shows why follow-up for so many years is crucial for understanding a condition such as dementia, which starts to develop many years before obvious symptoms appear or a diagnosis is made. The higher risk of dementia in women who did not participate in cognitive and social activities during the first 10 years of follow-up but not thereafter suggests that the gradual, insidious onset of dementia causes non-participation, rather than that non-participation causes dementia.

Relative risk of dementia, comparing the risk in women who did not participate in cognitive and social activities versus those who did[210]



Follow-up period	Relative risk (95% confidence interval)
0-4 years	1.41 (1.14 to 1.75)
5-9 years	1.21 (1.11 to 1.32)
10+ years	0.99 (0.95 to 1.03)

210 Figure adapted from Floud S et al. *Cognitive and social activities and long-term dementia risk: the prospective UK Million Women Study.* Lancet Public Health 2021 (https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(20)30284-X/fulltext).

If a person has consented to take part, engaged and invested time in a research study, this commitment should be respected and valued.[211] Indeed, the existence of ethically approved consent and active engagement in a study is a clear indication of a research participant's desire and expectation for their linked data to be used in research for the public good.

Well-designed research studies, particularly those aiming to support a wide range of research questions, invest considerable time and thought in ensuring that their consent processes are as robust, explicit and durable as possible. For example, participants may be invited to provide their consent to research uses by many different types of research organisations, including universities, charities, and commercial companies. However, the long-term nature of these studies means both that it may not be possible to predict in advance all potential beneficial future uses and users of research participants' data and that even well-designed participant information and consent materials may become dated as ethical and societal expectations change over time.

A challenge that may be faced by longitudinal studies is the assessment of the validity and robustness of consent obtained many years earlier, since which time scientific opportunities, information technologies and relevant legislation, regulation and public attitudes may well have changed. Any assessment of 'consent' therefore needs a flexible and proportionate approach. It should be undertaken with input from those with a clear understanding of the research value, aims and requirements of longitudinal studies, the practical difficulties of seeking updated or refreshed consent from research participants years after their recruitment, and the relevant ethical, legal and regulatory frameworks. It should also be informed by participant representatives to ensure public and participant views are accounted for in these deliberations.

Further, participants may change their minds during follow-up (although this is very rare in most longitudinal studies) and their wishes must be respected. Seeking updated or refreshed consent from all surviving participants periodically (for example every few years) during follow-up may seem an attractive option. However, it is often impractical or impossible to recontact and obtain responses from all research participants in a longitudinal study, especially when the study is large and participants are widely geographically distributed.[212] Non-response to consent requests can introduce significant bias and may make research findings less generalisable across different types of people and communities. Most research studies handle these challenges through providing participants with information and seeking explicit consent at recruitment, and then maintaining consent through regular newsletter and website updates and providing clear routes to withdraw from the study at any time during follow-up if participants wish to.

Legal and ethical principles allow the use of health data beyond the initial consent when this is in the public interest and measures are in place to protect participants' interests. The mechanisms for this vary between England and Wales, Scotland and Northern Ireland. In England, Section 251 of the National Health Service Act 2006 and the COPI Regulations provide a mechanism (see also section 3.2.3).

211 Unfortunately, longitudinal cohort studies have not always been able to honour the consent and expectations of their research participants. E.g., 15 or more years after their recruitment to UK Biobank, linkage of general practice records for all half million participants in the UK Biobank study to support the full range of research has still not yet occurred, despite the explicit consent of all the participants (see https://www.ukbiobank.ac.uk/using-gp-data-of-uk-biobank-participants). This issue is also discussed in sections 2.2, 2.3, 3.1.2, 6.3 and 7.2.1.

212 E.g., in 2013, the Avon Longitudinal Study of Parents and Children recontacted participants, originally recruited in the 1990s, to seek their consent for various types of data linkage. From among over 13,000 participants contacted (by post), only just over a quarter responded. Of the responders, 85–95% provided their consent for a wide range of linkages to health, education, economic and criminal justice data (see https://www.closer.ac.uk/wp-content/uploads/Boyd_CLOSER_20130702.pdf).

The successful management of consent and the application of these alternative approaches has varied considerably, with limited success.[213] One initiative that is successfully addressing several of the challenges is the Longitudinal Linkage Collaboration, which provides an efficient and coordinated mechanism to support the data linkage requirements of multiple UK population-based longitudinal studies.[214] An initiative led by Health Data Research UK's BHF Data Science Centre seeks to provide similar support for disease-based cohorts, initially focusing on cardiovascular cohorts but aiming to expand the service for all relevant disease-based cohorts across the UK.[215]

3.3.4 Research readiness registers

The UK has several registers of people who have signed up to note their interest in being invited to take part in medical research studies, often following an initial contact with healthcare services. Some are focused on research on a specific disease (for example Join Dementia Research for dementia[216]), but most are for research in any health area.[217] These registers maintain a database of contacts of the potential volunteers, including electronic (email or mobile phone) contacts, so that invitations to consider taking part in a research study can be sent to large numbers of people at low cost. For some research studies, these registers can help to boost or accelerate recruitment. However, at the time of writing, only a few hundred thousand people were signed up to these registers, a tiny fraction of the population of the UK. Because they are people who have expressed their enthusiasm to take part in research, the proportion responding positively to any invitation to take part in a study is likely to be higher than would be the case for unselected people from the wider population. But recruitment via services such as NHS DigiTrials, which can issue invitations to anyone based on age, sex, geographic or health characteristics from the whole population of England (57 million people), can invite people from a much larger pool of potential participants. This mechanism has underpinned large-scale recruitment to very large clinical trials such as the NHS-Galleri cancer detection trial (which recruited 140,000 people), or observational studies such as Our Future Health (which aims to recruit five million adult volunteers from across the UK).[218] Some, but by no means all, studies depend on recruitment by healthcare professionals in healthcare settings – for example a clinical trial of treatment for acute stroke, which needs to recruit people with suspected stroke immediately after their presentation in the hospital emergency department.

213 See Boyd A. *Understanding population data for inclusive longitudinal research*, a report for the Econoic and Social Research Council, 2021 (https://www.ukri.org/wp-content/uploads/2021/12/ESRC-240322-Understanding-Population-Data-for-Inclusive-longitudinal-research-V2.pdf).

214 See https://ukllc.ac.uk/.

215 See https://bhfdatasciencecentre.org/areas/cohorts/.

216 See https://www.joindementiaresearch.nihr.ac.uk/.

217 E.g. see https://bepartofresearch.nihr.ac.uk/, https://www.registerforshare.org/ and https://bepartofresearch.nihr.ac.uk/taking-part/uk-research-registries/index.

218 See https://digital.nhs.uk/services/nhs-digitrials.

3.4 Health-relevant data generated through environmental monitoring

3.4.1 Sources of environmental monitoring data

Climate, weather, air and noise pollution, together with the urban and rural buildings and landscapes that we live and work in, are among the most important determinants of health and health inequalities. Understanding how and why these influence our health and wellbeing is crucial for informing public health and healthcare policy.[219] Other areas of policy also need the best possible understanding of environmental health impacts. These include the response to the health impacts of climate change and extreme weather events; strategies on flood defences and air pollution monitoring; planning towns, cities, transport, work and green spaces; and traffic and building regulations.

Discussed in this section are several sources of data highly relevant to our health that can generate understanding about the impact of the environment on health.

Data from monitoring the weather and climate

Climate and weather conditions (for example temperature, rainfall, humidity, sun exposure and wind) affect people's health in ways that are incompletely understood. The UK Met Office has both active and archived data resources of major interest in this area, along with considerable expertise in their use, including for studies of the impact of weather and climate on health.[220] The Met Office Integrated Data Archive System (MIDAS) includes quality-controlled data, updated daily, on multiple meteorological variables (for example temperature, rainfall, wind speed, pollen count) from hundreds of weather stations across the UK. Linkage to national health data sources is not yet established to support the full range of potential beneficial uses. However, in 2020, the UKHSA established its Environmental Health Surveillance System (EPHSS). This builds on earlier developments, including a research council-funded initiative to connect diverse databases to improve understanding of the links between climate, environment, and human health. The EPHSS includes a Met Office data interface, enabling access to meteorological data that can be linked for public health purposes to health data held by UKHSA. Although currently only available for internal UKHSA access, there are plans for this to become externally accessible.[221]

As investment in research in this area gathers pace, researcher access to well-curated data linking weather and climate data to a wide range of health-related data at population-wide scale will be essential for rapid progress and societal impact. Alignment between new publicly funded initiatives such as the UKHSA's recently launched Centre for Climate Change

219 See UK Health Security Agency report *Preparedness for environmental hazards 2023:* https://assets.publishing.service.gov.uk/media/646b556da726f6000ccebb80/UKHSA_Advisory_Board_-_Preparedness_for_Environmental_Hazards.pdf.

220 See https://www.metoffice.gov.uk/research/applied/science-health-strategy.

221 See https://www.gov.uk/government/publications/environmental-public-health-surveillance-system.

and Health Security[222] and the UK Research and Innovation funded Centre for Climate Change and Health[223] will also be critical.

Data generated through air pollution monitoring

Air pollution has adverse effects on a range of health conditions, including respiratory diseases, heart disease, stroke, mental health and others. Analyses of existing data should improve our understanding of which pollutants are of most concern, how their effects are mediated and who is most at risk. This understanding in turn should inform policies to reduce this risk and to address the inequalities arising from its uneven distribution across the population. Several hundred UK-wide national monitoring sites monitor a wide range of air pollutants on behalf of the Department for Environment Food and Rural Affairs (DEFRA) and the devolved administrations. Large amounts of measurement data on these pollutants can be downloaded from DEFRA's Air Information Resource[224] and, with appropriate methodological expertise, can be linked via location information to a wide range of national health data sources (see following sections).

Data on the built environment

Our built environment includes road layout, housing density, availability of open or green spaces and location of health promoting and inhibiting facilities (such as sports centres and fast-food outlets). These are increasingly recognised to have an impact on our health, since they influence people's physical activity, lifestyle, social interactions and general wellbeing, affecting health outcomes such as obesity, mental health, heart disease, stroke and diabetes. Specialist methods can be used to generate location-based measures of the built environment using a range of UK-wide spatial data,[225] for example UK Ordnance Survey topography, transport network and address data.[226] These can then be linked using location information to person-level health data sources at national scale.

Data on noise and traffic

A growing body of evidence shows that major sources of noise such as road traffic, rail and aircraft affect health and wellbeing.[227] National data are generated through five yearly national strategic noise mapping exercises.[228] Similar to weather, climate and air pollution sources, specialist location-based linkage methods could be used to link these to national health data, enabling analyses to investigate the potential impact on a range of health outcomes.

Data on environmental radiation exposure

Exposure to radiation can cause health problems, including acute effects such as radiation sickness, and longer-term effects such as cancer. Relevant sources of radiation include naturally occurring radiation in the environment, radioactivity discharged into the environment by human processes, medical use of radiation, radiation used in industry, and radiation in items used and consumed by members of the public.

222 See UKHSA Report, *Health Effects of Climate Change in the UK: State of the evidence 2023*: https://assets.publishing.service.gov.uk/media/659ff6a93308d200131fbe78/HECC-report-2023-overview.pdf.

223 See https://www.ukri.org/opportunity/centre-in-climate-change-and-health/.

224 See https://uk-air.defra.gov.uk/air-pollution/.

225 E.g. see https://www.tandfonline.com/doi/full/10.1080/19475683.2015.1027791.

226 See https://www.ordnancesurvey.co.uk/products.

227 See https://ukhsa.blog.gov.uk/2023/06/29/noise-pollution-mapping-the-health-impacts-of-transportation-noise-in-england/.

228 E.g. see: https://www.data.gov.uk/dataset/d461bbc1-eb51-4852-8a9a-45dbf28aa230/noise-exposure-data-round-3; https://noise.environment.gov.scot/noise-statistics.html; https://datamap.gov.wales/layergroups/geonode:Environmental_Noise_Mapping_2022; https://www.daera-ni.gov.uk/services/noise-maps.

Monitoring of various environmental sources of radiation is coordinated on behalf of national UK and devolved governments by various national agencies, including the UK Environment Agency, UK Health Security Agency (UKHSA), UK Food Standards Agency, Scottish Environment Protection Agency, Natural Resources Wales and the Northern Ireland Environment Agency.[229]

Previous modelling by Public Health England (PHE), based on a range of data sources, found that average exposure to radiation in the UK from all sources of artificial radioactivity in food and the environment was well below legal safety limits. Most (85%) of this radiation exposure arises from ubiquitous radiation in the environment, much of it from radon.[230] No major changes in environmental radioactivity levels have been reported in the last two decades since the PHE modelling exercise.[231] However, as part of its national role in informing health security policy, the UKHSA continues to measure radon in UK homes, to routinely survey medical use of radiation in UK healthcare settings and to provide advice on occupational radiation monitoring more broadly.

There is no routine linkage of the various sources of radiation monitoring data to routinely collected healthcare data on health outcomes. However, this would in theory be possible using a combination of person-level and location-based linkage techniques and could be very helpful to support policy relevant research into the effects on health of various types of radiation exposure.

3.4.2 Key issues in the use and linkage of these data

It is beyond the scope of this review to explore in detail the full range of relevant environmental exposure data sources. However, there are some key points to highlight about how environmental data like these can and should be used to transform our understanding of health and wellbeing:

Value in linking environmental monitoring data with health data

The real value in these data comes from linking them at population-wide scale to health data. For example, linking data on air pollution measures in different locations to data about the health of people living and/or working in those locations enables studies of the impact of air pollution on health outcomes such as asthma and heart disease.

Linkage at location as well as at person level

For most sources of health-relevant data discussed up to this point, we have referred to person-level linkage of data sources – where different sources of data are linked together for each individual person. But environmental exposures depend on a person's location. This means that linkage of different data sources based on where a person is (or has been) may be crucial. Fortunately, many national, regional and local environmental data sources include location information, which can be used to link relevant environmental exposure measures to a person (or people) based on their postcode or the address of their home, school, work or other relevant locations. Using location-based information for data linkage and analysis can generate insights of great benefit for the public's health, but robust

229 See https://www.gov.uk/guidance/monitoring-radioactivity.

230 See https://assets.publishing.service.gov.uk/media/5a818b0b40f0b62305b8f855/PHE-CRCE-026_-_V1-1.pdf.

231 See https://www.gov.uk/government/publications/radioactivity-in-food-and-the-environment-rife-reports/rife-28-summary-radioactivity-in-food-and-the-environment-2022.

mechanisms must be used to protect the privacy of the people whose data are being linked and analysed. Chapter 4 discusses the linkage of different sources of data in more detail.

Level, timing, duration and pattern of environmental exposures

The impact of environmental exposures on different health conditions does not depend only on whether someone has been exposed or not. The level of exposure, for example intensity of noise or density of air pollutants, may be critical. The timing of the exposure, for example the person's age, or the precise time and date of exposure may also be important. Examples that illustrate this include:

- Infants and the elderly may be especially vulnerable to extremes of temperature and humidity.
- Daily, weekly, seasonal or year-on-year variation in environmental exposures (such as weather) and in people's locations (for example home, work or on holiday) means that timing of exposure can affect their impact.
- The health impacts of some exposures may occur quickly (for example an increase in acute asthma attacks triggered by a sudden reduction in air quality), while for others the effects may not appear until many years later (for example the skin cancer, melanoma, may be associated with excessive sun exposure many years earlier).
- The duration and pattern of an environmental exposure over time may also determine its health impact -for example, repeated and/or prolonged exposure to the sun's ultraviolet rays is likely to be important in the development of melanoma.

Appreciating that features such as the level, timing, duration and pattern of different environmental factors might influence our health will determine which data are useful and how they should be used. National environmental exposure records with UK-wide geographic coverage that go back for many years and include detailed measurements according to time and location, are likely to be particularly valuable. For example, properly investigating the impact of air pollution on a neurodegenerative condition that develops insidiously over many years before symptoms appear (for example motor neurone disease or Parkinson's disease) will benefit from detailed measurements of different types of air pollution by time and location, covering a period before disease onset of many years. Several national UK data sources (see section 3.4.1) have this sort of detail but have not yet been linked to the wide range of national health data that also exist, and then made accessible for analysis to fully realise the potential benefits.

Most environmental impacts are likely to be multifactorial

Finally, when it comes to understanding the causes of good and poor health, and of health inequalities, data from monitoring the environment will rarely if ever be the whole story. Most health problems do not have a single cause but are multifactorial: they are determined by a wide range of interacting factors, including our genes, lifestyle, diet, behaviours, social interactions and chance (or luck), as well the environmental factors discussed here. This emphasises the need for the linkage of environmental data resources to a wide range of other health data sources, so that the effects of all these different factors on our health can be better understood and addressed.

3.5 Health-relevant data generated by people

3.5.1 Data from personal electronic devices

Large volumes of data generated in our day-to-day lives through the electronic devices we interact with – for example our smartphones or smart watches – represent a potentially valuable source of information highly relevant to health. There is much to learn from these data, but in general we have barely scratched the surface, and gaining meaningful insights will not be easy.[232]

Around 9 out of 10 UK adults use a smartphone or wearable device (such as an Apple Watch or Fitbit). Smaller numbers use a medically approved device to monitor aspects of their health (such as heart rhythm or glucose levels). These generate a wide range of health-relevant data during people's daily lives, including about users' physical activity, sleep, environment, heart rate and rhythm, mood and more. Such information can be collected over months or years, giving a detailed picture of a how important measures of health change over time. Connecting data generated by people in this way to other health data, particularly from the NHS, has the potential to bring significant new understanding of the causes and consequences of health conditions such as arthritis, dementia, heart disease and mental health disorders, and – importantly – how these can be modified to benefit patients, families, carers and the wider population.

However, no large-scale initiative has yet established linkage of these types of data, collected over prolonged periods, to data on major health outcomes. This means that the large volume of potentially highly valuable information on the health of people across the UK is not yet being used as it could be to generate rich insights that inform understanding of health and disease, clinical practice and public health policy. We need national-scale initiatives where large numbers of smartphone and wearable device users are invited to consent to the linkage of their device data to routinely collected healthcare information to support innovative research. The linkage of data from smartphone and wearable devices of already well-characterised research participants into large population cohort studies with genomic data, such UK Biobank and Our Future Health, is another promising route to addressing questions such as how people's changing patterns of physical activity and heart rhythm might influence the development of health conditions, such as neurodegenerative diseases, years later.

Realising this opportunity will involve overcoming several challenges. These include:

- addressing 'digital inequality' (i.e., not all people own or use a smartphone or wearable device) to avoid biasing studies or worsening health inequalities;
- developing secure, robust methods for collecting and harmonising similar data items from different devices, and for storing, accessing and analysing these complex data;
- forging acceptable, transparent and productive partnerships with the companies who make and market the relevant devices;
- putting patients and the public at the heart of setting priorities for using these data, and advising on approaches to earn trust and gain consent to link data from people's devices to their routinely collected health data.

232 See Dixon W et al. Charting a course for smartphones and wearables to transform population health research. J Med Internet Res 2023. https://www.jmir.org/2023/1/e42449/.

3.5.2 Consumer loyalty card data

Each of us generates large quantities of data when we buy things, whether online or physically in various retail outlets.[233] There is increasing interest in the value of these data for understanding the association between how much we spend, the things we buy and our health. This could be addressed by linking data from consumer loyalty card databases (for example from supermarkets such as Tesco or pharmacies such as Boots) to routinely collected national health data. As with data from smartphones and wearables, there are significant potential benefits but also challenges. An illustrative example is provided by the UK-based Cancer Loyalty Card Study (Box 3.13).[234]

Box 3.13 Could shopping data from consumer loyalty cards help to detect ovarian cancer earlier?

Earlier diagnosis is urgently needed to improve the outcomes of ovarian cancer, particularly survival. In the Cancer Loyalty Card Study (CLOCS), researchers are investigating whether commercial data on women's shopping behaviours, collected through loyalty card use at UK high street retailers, might provide early warning signs of ovarian cancer. Early results are promising, showing that there may be an increase in the purchases of painkillers and indigestion tablets up to 10–12 months before diagnosis. This is encouraging for the expansion of this approach to address other relevant health questions, broadening the benefits for patients and the public. The CLOCS research team is working with patients and members of the public to develop partnerships with relevant companies, design processes for recruitment and consent, tackle legal and regulatory data sharing issues, and establish data acquisition, storage, access and analysis methods.

233 E.g. see https://www.cdrc.ac.uk/ and https://www.ons.gov.uk/economy/inflationandpriceindices/articles/introducingalternativedatasourcesintoconsumerpricestatistics/april2022#alternative-data-sources.

234 See https://www.clocsproject.org.uk/publications.

Chapter 4

The power of linking different sources of data

In this chapter

4.1 Benefits of linking data

Different sources of health-relevant data are informative on their own for specific purposes and this is most often how they are used. However, it is when different data sources are linked together – especially at whole-population scale – that truly transformational insights start to emerge.

4.1.1 Illustrating the benefits

Generating insights that would be impossible from any single data source

Many valuable data-driven discoveries that can inform and drive improvements in healthcare and policy would simply not be possible through analysis of any single data source. Several examples are shown in Box 4.1.

Box 4.1 Benefits of linking data from different sources

Linking screening data to cancer and death registry data to investigate which breast cancer screening strategies work best

ATHENA-M (Observational study of Age, test THreshold and frequency on English NAtional Mammography screening outcomes) links data from routine breast screening records (screening invitations, attendances and test results) from 1988 to 2018 to national cancer registry (breast cancer diagnoses and treatment) and national mortality data (date and cause of death).[235] The database is the most complete set of English breast screening records and outcomes ever created. It includes information on over 11 million women invited for screening and followed for an average of over 12 years. Researchers are using the data to evaluate and optimise the breast screening programme in England, answering questions such as:

- How often should women be invited to breast cancer screening?
- What is the most appropriate age range to offer breast cancer screening?
- What types of abnormalities on mammograms should be investigated further?

Linking healthcare data to data from other sectors to explore the relationships between health, education and social care

ECHILD (Education and Child Health Insights from Linked Data) links data from NHS England (Hospital Episode Statistics, Mental Health Services Data Set and Maternity Services Data Set), the Department of Education (National Pupil Database with information on pupil and school characteristics, educational attainment, social care), and the Office for National Statistics death register.[236] These linked data create a longitudinal database that follows the lives of around 20 million children and young people in England born since 1984. Researchers are using it to explore the relationships between health, education and social care from childhood to adulthood, generating insights to inform policy and practice. Researchers are using these data to answer questions such as:

- Do school absences explain the association between chronic ill health and lower school attainment?
- What are the health outcomes in young adults who had contact with social care services or special educational needs in childhood?
- How does support for children with special educational needs and disability affect their health in later life?

Additional linkage to general practice data would greatly enhance the range of health outcomes that could be studied.

235 See Brettschneider J et al. *ATHENA-M*. Br J Radiol 2024 (https://academic.oup.com/bjr/article/97/1153/98/7470406).

236 See ECHILD: https://www.ucl.ac.uk/child-health/sites/child_health/files/echild_user_guide_v2.pdf.

Linking national heart failure audit data to hospital episodes and death registry data to assess the impact of specialist heart failure care on health outcomes

There are many disease- or domain-specific national audits and registries (section 3.2.12). Linking these to a range of other data sources (such as general practice records or hospital episodes data) can provide additional information on later health outcomes, previous medical diagnoses, prescribed medications and so on. Within the NHS England secure data environment for England, researchers have linked data from several specialist cardiovascular audits to many other health data sources. Linking the heart failure audit data to hospital and death data, providing information on health outcomes and causes of death, enabled England-wide analyses assessing the impact of specialist heart failure care on the health outcomes of heart failure patients. These showed that heart failure patients receiving specialist input in hospital had better health outcomes, even after accounting for the effects of age, sex, ethnicity, severity of heart failure, and other health conditions.[237]

237 See Cannata A et al. *A nationwide, population-based study on specialized care for acute heart failure throughout the COVID-19 pandemic.* Eur J Heart Failure 2024 (https://onlinelibrary.wiley.com/doi/10.1002/ejhf.3306).

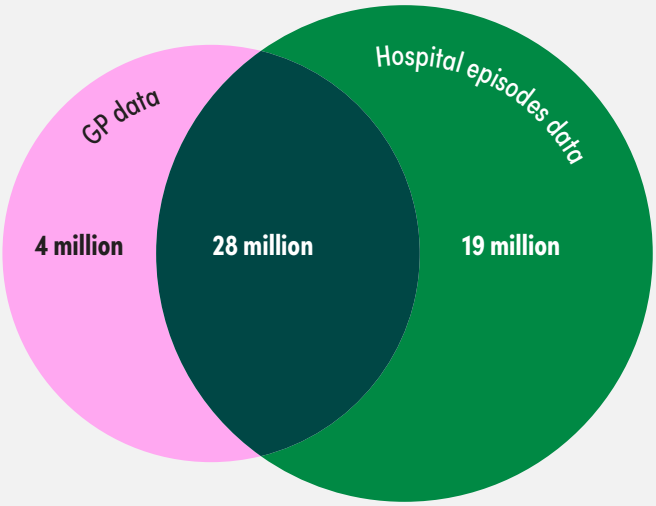
Building a more complete and accurate picture of people's characteristics

There are other ways in which linking data from different sources can bring major benefits to patients and the wider population. For example, as illustrated in Box 4.2, linkage between data sources provides a more complete and accurate picture of each person represented in the data, and can help to identify where information gaps exist despite the linkage and work out how best to fill them. This reduces the possibility of inaccurate or under-representation of certain subgroups (for example those from more deprived backgrounds or ethnic minorities).

Box 4.2 Linking whole-population general practice and hospital data to improve completeness of ethnicity information[238]

Most health data analyses need information on key characteristics of the people represented in their data, including their ethnicity. Unfortunately, this information is often incomplete in routinely collected health data. In 2020, data from several different healthcare sources were available for a total population of around 54 million people in the NHS England secure data environment. Data on ethnicity was available for around 32 million people (59% of the total) from their general practice records and from around 47 million people (87% of the total) from their hospital records. By combining both sources of data, ethnicity information was available on 51 million people (almost 95% of the total), showing how linking different sources of data together provides a more complete picture than any single source alone.

Number of people with ethnicity data in GP and hospital records in the NHS England secure data environment in 2020



238 See Wood A et al. *Linked electronic health records for research on a nationwide cohort of more than 54 million people in England.* British Medical Journal 2021 (https://www.bmj.com/content/373/bmj.n826).

Linking different data sources to find all people with a specific health condition

As well as plugging information gaps about people's characteristics, such as their ethnicity, linking data from different sources can help to ensure that as many people as possible with a particular health condition are included in any analysis of that condition. Box 4.3 gives an example of how linking data from multiple sources, including general practices and hospitals, helps ascertain a much larger number and wider range of severity of cases for many different health conditions than would be found through any single data source.

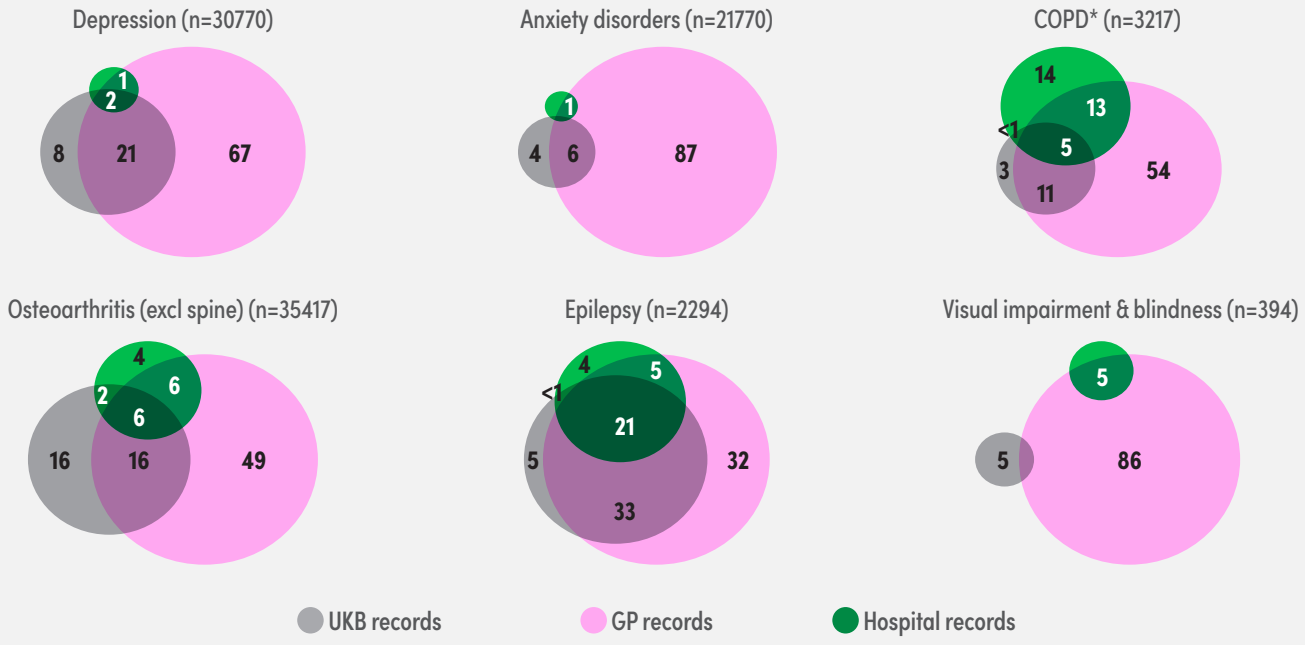
Box 4.3 Linking different sources of data in UK Biobank to improve research into different health conditions

To generate robust findings about specific health conditions, researchers try to analyse data from as many people with the condition as possible. Linking different sources of data helps them to find the relevant participants.

For instance, an analysis of data from around 170,000 UK Biobank participants, aged 40–69 years at recruitment, looked at which participants had one or more of 80 long-term health conditions when recruited. The data sources were general practice records, hospital records and UK Biobank records of participant self-reported health conditions.

Combining data from all three data sources, 85% of participants had at least one of the 80 health conditions. For 62 of the 80 conditions, general practice data was the source that identified the largest proportion of people with the condition. The Venn diagrams below show the sources of data identifying participants with six of these conditions. They show the importance of using multiple different sources of data for research on specific health conditions. They also highlight the relevance of general practice data for identifying people with health conditions that have not resulted in admission to hospital and so would not be identified within hospital records.

Combining data from different sources identifies more people with different health conditions (numbers in the circles are %).[239]



239 Figure adapted from Prigge et al. *Robustly Measuring Multiple Long-Term Health Conditions Using Disparate Linked Datasets in UK Biobank* (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4863974).

*COPD: chronic obstructive pulmonary disease

Linkage to health records to follow the health of research volunteers

Efficient, comprehensive follow-up in some large-scale observational cohorts and randomised clinical trials has been revolutionised by the ability to link to data about research participants within national health data sources. Box 4.4 shows an example of this type of follow-up.

Researchers generally do this linkage with the explicit consent of the research participants. It can be used to supplement or replace more conventional methods of follow-up in research studies, such as inviting research participants to attend follow-up clinics, or contacting patients via telephone or email to ask them relevant questions about their health.

Box 4.4 Linking to national health data for trial participant follow-up to improve efficiency and reduce costs

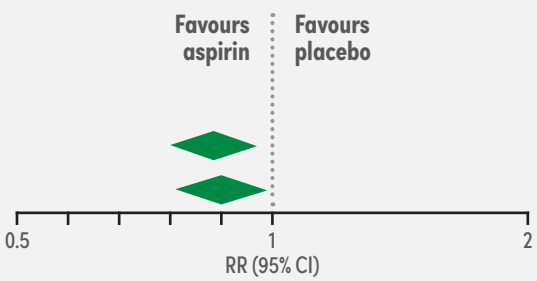
Routinely collected data could greatly increase efficiency and reduce the cost of conducting trials. The ASCEND randomised clinical trial compared aspirin with placebo for preventing serious vascular events (heart attacks, strokes, 'revascularisation' procedures to open narrowed blood vessels, or death from vascular disease) in 15,500 people with diabetes in the UK.

Researchers used two methods to follow up participants. The first involved contacting participants to ask whether they had had a heart attack, stroke or revascularisation procedure.

Clinical experts then reviewed clinical records to verify each report ('adjudicated direct follow-up'). The second approach identified the relevant health outcomes by linking to national hospital data and the national death register ('routine data follow-up'). The two approaches produced almost identical trial results. This suggests that routinely collected hospital and death registry data in the UK could replace more costly methods of follow-up for some types of outcomes – in this case serious vascular events – in randomised trials.

Comparison of different follow-up approaches in the ASCEND trial showing similar rate ratios (aspirin versus placebo) for serious vascular events[240]

	Treatment, No. (%)		
Outcome	Aspirin (n=7740)	Placebo (n=7740)	RR (95% CI)
Any serious vascular event or revascularization			
Adjudicated direct follow-up	833 (10.8)	936 (12.1)	0.88 (0.80-0.97)
Routine data follow-up	726 (9.4)	805 (10.4)	0.90 (0.81-0.99)



240 Figure adapted from Harper C et al. *Comparison of the Accuracy and Completeness of Records of Serious Vascular Events in Routinely Collected Data vs Clinical Trial-Adjudicated Direct Follow-up Data in the UK.* JAMA Netwk Open 2021 (https://pubmed.ncbi.nlm.nih.gov/34962561/).

4.2 How is data linkage done?

4.2.1 Background to linkage methods

Linking at person level between different sources of data involves identifying where the same people are represented within two or more datasets and joining the data together for each of these people. **Matching** is the process of identifying the same person across different data sources, while **linkage** is the joining together of the data sources. Person-level matching can be based on finding the same unique identifier or combination of identifiers across more than one data source.

The most straightforward approach to linkage at the person level occurs via 'deterministic' linkage, where the data sources to be linked contain the same single uniquely identifying data item for every person in each data source. The NHS number in England, Wales and Northern Ireland, and the CHI number in Scotland, are examples of unique person identifiers. However, it is rarely this straightforward. For example, taking NHS number as an example:

- a few people have more than one NHS number;[241]
- NHS numbers in some systems may contain errors[242] and so fail to generate matches when they should.

Where a unique identifier matches across datasets, deterministic matching and linkage can occur. Most – but not all – NHS data sources now include either the NHS number or Scottish CHI number. But where this is not the case, either for an entire data source or because the information is missing for some records in that source, combinations of other identifiers may be used for matching and linkage.[243] For example, deterministic matching and linkage may occur when each person's forename, surname, date of birth and postcode of residence all match for the records across the data sources to be linked. When linkage is based on some, but not all, of a combination of identifiers matching,[244] it is called probabilistic linkage, because the confidence in the accurate linkage of each record is not 100%. Sets of rules for deciding the number and type of matching identifiers needed to link records accurately (with high, albeit not 100%, confidence) between different data sources have been developed by many organisations that link different sources of data.

241 E.g., if someone is treated as an emergency in hospital while unconscious or otherwise unable to provide their identifying details, they may be issued with a temporary new NHS number; in the past, some people may have been issued with a new NHS number when they moved and registered with a new general practice but could not recall their previous practice details or NHS number.

242 This is likely to have happened more in the past when NHS numbers were more commonly entered into systems manually, whereas nowadays this process is more likely to be automated.

243 See https://www.adruk.org/learning-hub/skills-and-resources-to-use-administrative-data/navigating-administrative-data/#c8816.

244 E.g., this could occur if a person's unmarried surname is used in one source, and a different, married surname is used in another; or if a person's name is mis-spelt in one or more source.

Linking data sources, which are owned or controlled by different organisations, needs particular security considerations. Processes need to protect the privacy of the individuals represented in the data and to minimise the sharing of identifiable information between organisations. In many situations, linkage of data sources held by different organisations is conducted by a so-called 'trusted third-party' organisation. Simple, yet robust and secure, systems using data encryption interfaces to ensure that no directly identifying information passes between organisations may also be used.[245] Sometimes more sophisticated methods or software might be needed for more complex data linkage scenarios.[246]

Finally, as discussed earlier, location-based information, for example postcode or household, is needed to link data on a wide range of environmental exposures to people and groups of people, based on their current and previous work, home, education or other locations. Specialised linkage methods based on location reference information (be it unique property reference numbers, unique street reference numbers, postcodes or something else),[247] enable the linkage of data sources at place – as well as person – level. Such linkages require careful assessment and management of data security and privacy, but they bring substantial benefits. For example, they underpin analyses of how and why measurable household characteristics (such as space, overcrowding, damp, mould, radon exposure) and location-based exposures (such as air pollution, noise, and weather conditions) impact the health and wellbeing of different subgroups (for example rich and poor, children, adults and elderly, different ethnicities) across society.

4.2.2 Linkage approaches of relevant national organisations

The Office for National Statistics (ONS) provides detailed methodological guidance and national coordination in linking data across government departments.[248] These approaches, and further emerging linkage methods, will be an important part of the services provided by the ONS's evolving Integrated Data Service. This aims to bring together diverse data from across a range of government departments and other sources to enable secure access for faster and wider collaborative analysis for public benefit.[249]

The systems and methods used by national health data custodians for linking data from different sources at person level have until recently been poorly described or even opaque. This has limited the validity and reliability of the insights from national linked data. However, NHS England has recently provided detailed descriptions of how its Personal Demographic and Master Person Services are used to facilitate secure, accurate linkage between different health data sources covering the entire population of England.[250] Provision of this important methodological detail is welcome. An essential next step will be for NHS England to provide record-level information on matching and linkage quality for researchers and analysts working with NHS England linked data. This is important because record match quality may vary across population subgroups. For example, match quality may be lower among people from ethnic minority groups. The ability to assess, account for and – in due course – correct variations in match quality by key characteristics

245 E.g. see https://www.openpseudonymiser.org/.

246 https://datasciencecampus.ons.gov.uk/developing-a-privacy-preserving-record-linkage-toolkit/.

247 E.g. see https://www.gov.uk/government/publications/open-standards-for-government/identifying-property-and-street-information.

248 See https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods.

249 See https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice/integrateddataservice.

250 See https://digital.nhs.uk/services/personal-demographics-service/master-person-service/the-person_id-handbook.

will prevent inequalities being embedded in the data used to generate insights that inform health policy.

We could not find readily accessible information in the public domain on the methodological details (for example decision rules, algorithms and statistical methods) of data linkages conducted by and on behalf of national organisations in the devolved nations. However, there is helpful information on the data linkage principles and their implementation.

In Scotland, the underlying basis of linkages between many NHS Scotland and non-NHS health-relevant data sources is the community health index (CHI). Like NHS England's Personal Demographic Service, this is a register of all users of NHS Scotland services, in which each person has a unique 10-digit CHI number.[251] Deterministic linkage between datasets across Scotland depends on the CHI, although, as in other parts of the UK, probabilistic methods are required when unique identifiers (the CHI in particular) are unavailable or missing. The Scottish Government's guiding principles for data linkages, based on several decades of experience in linking across different sources of data in Scotland, are generally relevant across the UK.[252] The Welsh SAIL Databank, hosted by Swansea University, working in partnership with Digital Health and Care Wales (DHCW), provides information on the principles and processes for linking datasets at the person and household level. As for the Scottish guiding principles, emphasis is placed on the separation of data linkage and data analysis functions through use of a trusted third party (DHCW).[253] Similar, albeit less detailed, information on linkage processes is available via Northern Ireland's Honest Broker Service.[254]

251 See https://publichealthscotland.scot/services/chi-linkage-and-indexing-chili/about-the-chi-linkage-and-indexing-team-chili/.

252 See https://www.gov.scot/publications/joined-up-data-better-decisions-guiding-principles-data-linkage/.

253 See https://saildatabank.com/governance/privacy-by-design/.

254 See https://bso.hscni.net/directorates/digital-operations/honest-broker-service/.

Chapter 5

Current and emerging routes of access to health-relevant data

In this chapter

5.1 Evolution of a network of national remotely accessible secure data environments

Notable developments in data platforms have boosted the ability of researchers, analysts and policymakers to use data to improve public health and patient care. The last few years have seen a rapid acceleration towards the wider use of secure data environments (SDEs) (Box 5.1) for research and analysis using health-related data, particularly in England.[255] However, such environments have been in use for access to and analysis of health and other administrative data sources for many years. They were established mainly to enable safe and secure access to de-identified data from health, care or other administrative systems for approved research, without each person necessarily providing their explicit consent. Access to data for approved purposes in the public interest is supported by different legal gateways (depending on the data source, type and country) combined with a social licence from the public.

255 See https://www.gov.uk/government/consultations/data-access-policy-update-proposed-draft/data-access-policy-update-proposed-draft.

Box 5.1 What is a secure data environment?[256]

A **secure data environment** (SDE) is a secure computing platform for storage of and remote access to data for analysis, together with the governance processes and people that enable such access. SDEs are increasingly used for accessing and analysing sensitive data such as health and social care data. Data within SDEs can only be accessed by approved analysts or researchers for specific approved purposes without the raw data ever leaving the environment. Users come to the data rather than the data going to them. Some people describe this as being like a reading library, where library books are not taken out of the library, in contrast to a lending library, where readers can take books out of the library. Uses of data within a SDE can be checked and audited, for example checking who analysed which data and when, and that this was in line with the specific approved use.

In most SDEs, analysts or researchers access person-level and record-level data that have been de-identified. This means the data have had all directly identifying information (such as names, addresses, NHS numbers and exact dates of birth) removed. Before analysts or researchers can export the results of their analyses (such as summary tables or graphs) out of the SDE, these must first be checked by trained output checkers (in some cases with the assistance of robust automated output checking processes) to ensure that they do not contain any information that might inadvertently lead to the identification of any person whose data are included in the analyses.

Well-designed SDEs improve the privacy and security of people's data. They can – and should – also improve the efficiency of research and analysis because the same carefully prepared data can be re-used for a wide range of analyses. SDEs can also promote collaboration between researchers and analysts based in many different organisations. SDEs in the UK must comply with the UK's robust legal and regulatory frameworks to keep data safe and ensure it is used correctly. SDEs holding sensitive data must also adhere to the Five Safes Framework, an internationally recognised system promoting the best practice in data security and privacy (see section 5.5).

Other commonly used terms that have the same meaning as secure data environment include:

- trusted research environment
- data safe haven
- secure processing environment
- research data library

256 For useful further information, see https://understandingpatientdata.org.uk/secure-data-environments.

Many years prior to the COVID-19 pandemic, national SDEs were established by national health data custodian bodies in Scotland and Wales. These enabled secure, remote access to de-identified data from a range of population-wide healthcare data sources for approved research projects. In both cases, these SDEs were established – and are provided – through close partnerships with universities, which have brought significant expertise in the design and management of systems for secure data storage, linkage, remote access and analysis, together with the ability to raise additional investment to supplement direct government funding.[257] In Northern Ireland, before the pandemic, the Health and Social Care Honest Broker Service provided approved researchers with secure access to some linked, de-identified health data, but this required physical attendance at the Safe Haven in Belfast and so was hardly used by researchers outside Northern Ireland. The Office for National Statistics (ONS) has a long-established SDE, the Secure Research Service, which has provided secure access to de-identified, linked population-wide administrative and survey data (covering England, England and Wales, or all UK nations, depending on the data source and type) for accredited researchers for over 15 years, albeit with limited inclusion of healthcare data derived from the NHS. There have been no reported data breaches over many years of researchers analysing de-identified record level data within these established SDEs.

Prior to the pandemic, there was no SDE for secure in-situ access to different sources of de-identified, linked, whole-population healthcare data from NHS England. Prior to 2020, NHS Digital (now part of NHS England) operated only a data dissemination (or data transfer) model of data access via its Data Access Request Service. Transfer from NHS Digital systems of health data from different sources was limited (certainly by comparison with demand, need and potential benefits), and linkage at whole-population scale of multiple different sources of data had never been conducted. The pressing needs of the pandemic changed this, driving several new initiatives to link and enable remote secure access to health data from an increasing range of different sources.

Some of the most prominent national-scale developments in England during the pandemic are described in Box 5.2.

257 E.g., through competitive grants from a wide range of research funding bodies.

Box 5.2 National-scale developments in England for secure access to health-relevant data during the pandemic

NHS England secure data environment (SDE)

A partnership between NHS Digital and the BHF Data Science Centre at Health Data Research UK led to the establishment of a remotely accessible SDE, hosted by NHS Digital. This was initially developed to support research on the cardiovascular drivers and consequences of COVID-19. For the first time, structured, coded data for the whole population of England (57 million people) from sources relevant to a very wide range of health conditions (including general practice, hospitals, community-dispensed medicines and registered deaths) were linked at person level and made securely available for approved research.[258] Over time, COVID-19 testing and vaccination data, together with an increasing range of specialist national datasets (for example specialist cardiovascular audits, and mental health and maternity services datasets) were also linked. With the support of the Government Chief Scientific Adviser's National Core Studies Data and Connectivity programme,[259] the BHF Data Science Centre initiative was extended to cover any approved COVID-19-related research (not just cardiovascular). In addition, tenancies were created for several other user groups within the SDE to extend its benefits, for example to support research on the cancer-related impacts of COVID-19 by the health data research hub, DATACAN, and to support analyses by the Department of Health and Social Care (DHSC). Having successfully supported many analyses generating policy relevant insights, the NHS Digital SDE has now been further developed to become the NHS England SDE, a major component of NHS England's evolving secure data access infrastructure.[260]

OpenSAFELY

OpenSAFELY is a secure, transparent, open-source software platform for analysis of electronic health records data. It was developed through a partnership between the Bennett Institute at the University of Oxford, the two main commercial general practice computer system suppliers for England, and NHS England.[261] The OpenSAFELY platform was initially established within the data centre of the computer system supplier The Phoenix Partnership (TPP), and subsequently deployed within the cloud data environment of England's other major general practice computer system supplier, Egton Medical Information Systems (EMIS). OpenSAFELY's deployment within these two systems enables secure access to detailed data from general practices covering almost the entire population of England. Rather than accessing patient record-level de-identified data, as occurs in most SDEs, approved researchers must instead write code against a synthetic dataset constructed to mimic the live data within the general practice computer systems. This code is then run on their behalf by a small team of developers working within these systems. The OpenSAFELY interface has been approved by the British Medical Association, Royal College of General Practitioners and the privacy group,

258 See https://bhfdatasciencecentre.org/areas/cvd-covid-uk-covid-impact/; Wood A et al. *Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource*. BMJ 2021 (https://www.bmj.com/content/373/bmj.n826); https://digital.nhs.uk/services/trusted-research-environment-service-for-england.

259 See https://www.hdruk.ac.uk/covid-19-data-and-connectivity/.

260 See https://digital.nhs.uk/services/secure-data-environment-service.

261 See https://www.opensafely.org/about/.

medConfidential, all organisations that are represented on the OpenSAFELY Oversight Board.[262] Linkage and access to data from a range of other sources (mainly NHS England but also several others), covering the population of England, are facilitated through transfer of de-identified data from these sources into the general practice computer systems.[263]

The ONS Public Health Data Asset

Supported by the Data and Connectivity National Core Study, the ONS and NHS England worked jointly to establish and make available the ONS Public Health Data Asset. This links data from the 2011 Census, the General Practice Extraction Service data for COVID-19 pandemic planning and research, hospital episode statistics for England and registered deaths.[264] The data can be accessed via the ONS Secure Research Service, The creation of this dataset demonstrated a mechanism for sharing of NHS England data with the ONS, essential for the future development of cross-sectoral linkages between health and non-health data for the English population.

262 See https://www.opensafely.org/governance/.

263 See https://docs.opensafely.org/data-sources/.

264 See Thomas S et al. *Study protocol for the use of time series forecasting and risk analyses to investigate the effect of the COVID-19 pandemic on hospital admissions associated with new-onset disability and frailty in a national, linked electronic health data setting.* BMJ Open 2023 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10201261/).

The three initiatives described in Box 5.2 represent significant advances, enabling secure access to health data at a scale, depth and breadth considered unimaginable prior to the pandemic. They have led to the generation of insights and research outputs with significant beneficial impact on clinical and public health practice and policy.[265] For example, they enabled analyses that demonstrated the effectiveness of COVID-19 vaccination across the population, or that showed reductions in prescribing and dispensing of medicines to prevent cardiovascular diseases during the early months of the COVID-19 pandemic (see Box 3.4). They have also produced advances in methods for the reproducible curation and analysis of the data they have linked and enabled access to,[266] as well as providing guidance and training in the use of these data for an increasing number of UK-based researchers and analysts (for example in the NHS, universities, charities, National Institute of Health and Care Excellence, the DHSC). By establishing all the necessary approvals for a wide range of research projects under broad programmes, and facilitating robust but efficient data access processes, the partnerships between NHS England and both OpenSAFELY and the BHF Data Science Centre have enabled much more rapid and efficient access to data for larger numbers of researchers and projects than could possibly have been handled directly through NHS England's Data Access Request Service working on a project-by-project basis.

Although motivated by the needs of COVID-19, these initiatives were established with the intention of creating advances in infrastructure for secure access to data that would endure beyond the pandemic. Unfortunately, several of the national health data assets that became available at England-wide scale for the first time during the pandemic still remain accessible only for COVID-19-specific analysis and research. This is because they were provided for these specific purposes only, under 'COPI notices' (see section 3.2.3) issued by the Secretary of State for Health and Social Care. NHS England's announcement in November 2023 of its plans to widen the use of the OpenSAFELY platform for research relevant to major non-COVID-19 conditions such as cancer, diabetes and asthma is welcome.[267] However, the necessary expansion of the existing legal direction to enable this has not yet occurred (as of 23 October 2024). And plans to extend the uses of those data currently available only for COVID-19-specific purposes within the NHS England national SDE, as well as to make additional datasets already held by NHS England accessible within this national SDE, face a range of difficulties. These include insufficient resources within the relevant teams in NHS England and unresolved data supply issues within relevant data provider organisations. For example, provision to NHS England of the several national cardiovascular audit datasets collected by the National Institute of Cardiac Outcomes Research is blocked due to what is referred to as an "external supplier issue".[268]

These new English population-wide SDE initiatives do not in themselves solve all data access challenges. DHSC data access policy developments mean that the data dissemination (or data transfer) route – as opposed to access within the data custodian's own SDE – is being used decreasingly for data-driven studies. This is especially so for analyses conducted without

265 See https://bhfdatasciencecentre.org/wp-content/uploads/2024/05/240229-CVD-COVID-UK-COVID-IMPACT-Research-Outputs.pdf; https://www.opensafely.org/research/#published.

266 See https://github.com/OpenSAFELY; https://github.com/bhfdsc.

267 See https://www.england.nhs.uk/2023/11/nhs-expands-use-of-secure-covid-19-research-platform-to-help-find-new-treatments-for-major-killer-conditions/.

268 See https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services/data-set-additions-and-improvements-report.

the explicit consent of research participants. However, the data dissemination route remains highly relevant and necessary for the secure transfer of linked health-relevant data to many observational and interventional research studies, including widely used research resources such as UK Biobank, and clinical trials (see sections 3.3.1 and 3.3.2). It is not realistic to expect NHS infrastructure to support the complex needs of these studies and their research communities. Indeed, many have developed or are developing their own specialist remote access SDEs to support researcher access. Neither do the new national English SDE initiatives yet have robust, transparent and scalable mechanisms for access to data for researchers from commercial companies, who continue to rely heavily on access to linked health data via the well-established Clinical Practice Research Datalink (CPRD), which covers about 30% of the English population (see section 5.3).

Funding for further development of the NHS England SDE is being made available through the NHS England national and regional Data for Research and Development programme. This has been running since April 2022 with £175 million over three years to fund the NHS Research SDE Network, including the NHS England SDE and 11 regional SDEs (section 5.2), as well as the NHS DigiTrials service (section 5.4.2) and national public deliberations on the use of health and care data.[269] Up to £8 million of additional funding over two years is being made available to extend the capability of OpenSAFELY to support 'beyond COVID-19' studies. Significant UK government funding has also been made available for the ONS's new Integrated Data Service which will in due course replace the Secure Research Service.

Investments from Administrative Data Research UK, Health Data Research UK (chiefly via National Core Studies Data and Connectivity funding), and directly from the governments of the UK and devolved administrations, prior to, during and since the pandemic, have facilitated further enhancements of the national SDEs in the devolved nations. In Scotland, the establishment of Research Data Scotland aims to provide durable national capability for linking healthcare to non-healthcare administrative data for research approved by Scotland's Health and Social Care and/or Statistics Public Benefit and Privacy Panel(s).[270] In Wales, the breadth of data available within the SAIL Databank, which already provided the most diverse healthcare and non-healthcare data anywhere in the UK prior to the pandemic, has been further increased. In Northern Ireland, a new, remotely accessible SDE, the Northern Ireland Trusted Research Environment (NITRE) for health and social care data, using similar technical infrastructure to the Welsh SAIL Databank, has been established through a partnership with Swansea University.[271] However, as in England, in Northern Ireland, a separate environment provides access for approved users to non-health and care data from other areas of government, making cross-sectoral linkages of health and care data to other health-relevant data sources more difficult.[272]

Figure 5.1 shows the national SDEs enabling access to data from the health and care systems across the four nations of the UK. Figure 5.2 provides more detail on the flows of data into, between and from the national secure data systems that support SDEs holding health and care data in England.

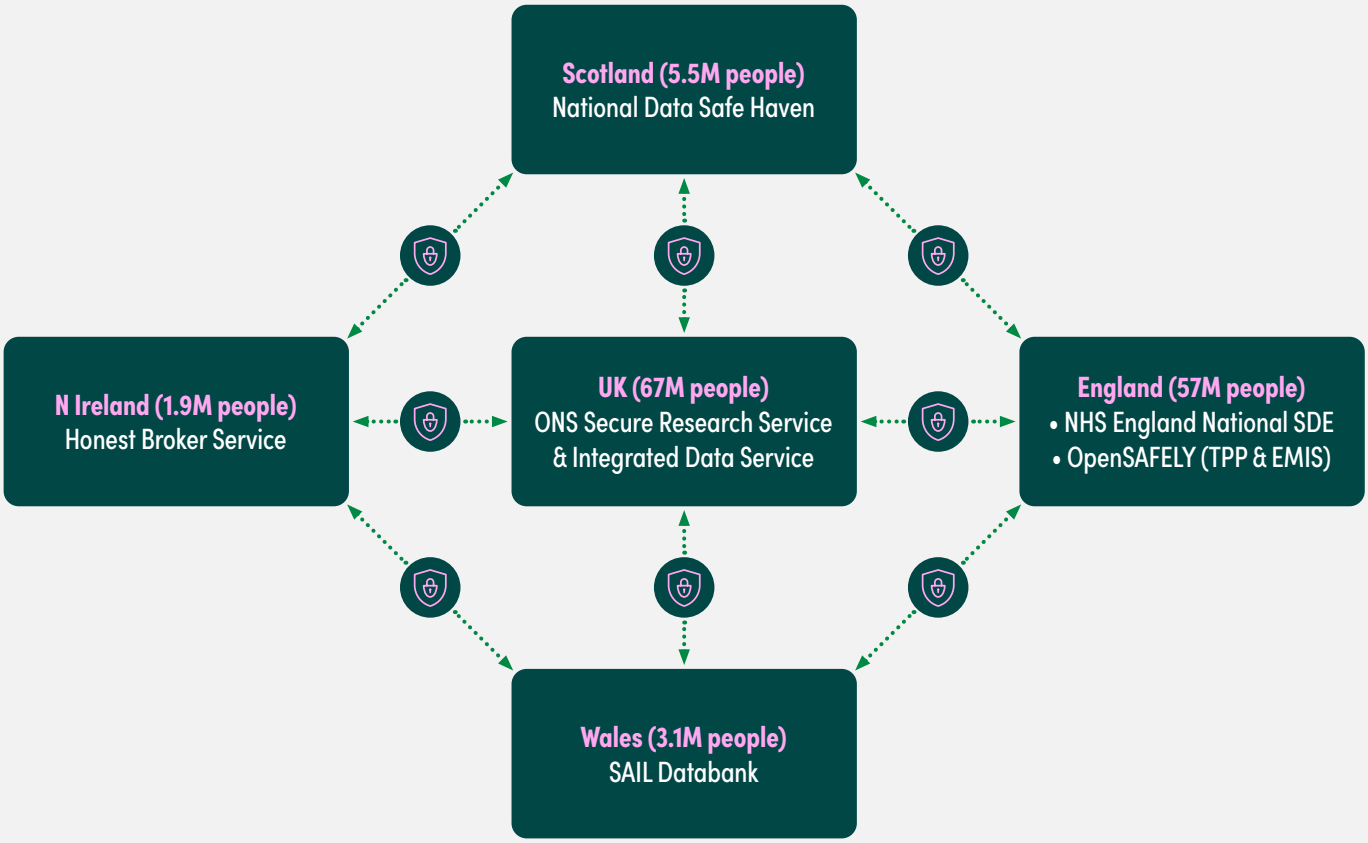
269 See https://transform.england.nhs.uk/key-tools-and-info/data-saves-lives/national-public-engagement-on-the-use-of-health-data/.

270 See https://www.researchdata.scot/.

271 See https://dhcni.hscni.net/digital-strategy/data/ and https://www.hdruk.ac.uk/organisations/northern-ireland-honest-broker-service-northern-ireland-health-and-social-care/.

272 See https://www.nisra.gov.uk/.

Figure 5.1 National secure data environments enabling access to whole-population data from the health and care systems across the four UK nations

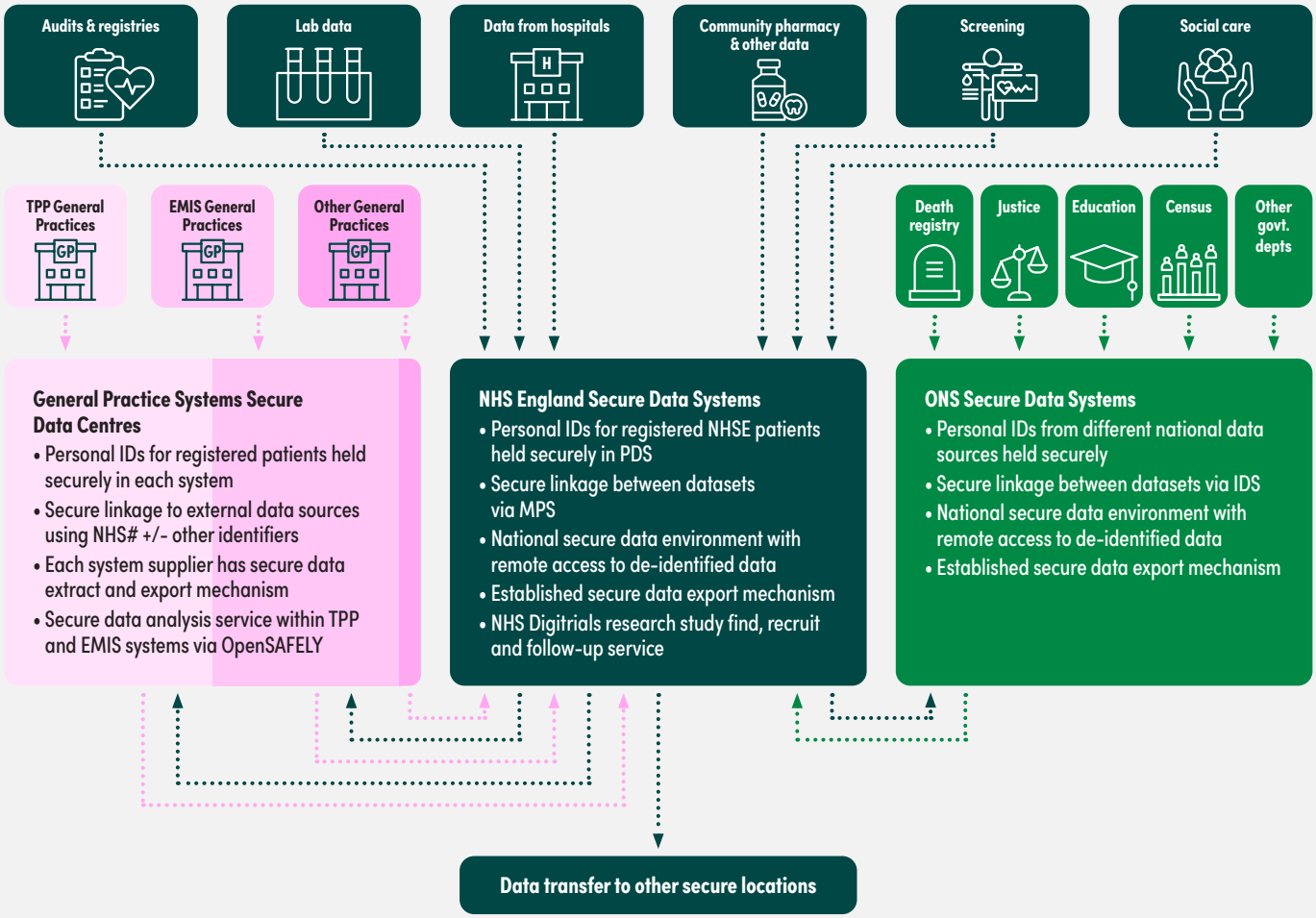


SDE: secure data environment; SAIL: secure anonymised information linkage; Open SAFELY (TPP & EMIS): OpenSAFELY operating within the data centres of England's two main primary care computer system suppliers, The Phoenix Partnership (TPP) and Egton Medical Information Systems (EMIS)

Capability for sharing data securely between national secure environments

Secure sharing of data (whether via secure data transfer or more complex methods such as 'federated queries') between these environments is possible but in practice very limited. Linkages of health and care data to other administrative data occur within a single national SDE in Wales (SAIL) and are also supported in Scotland's National Data Safe Haven.

Figure 5.2 Data flows into, between and from national, whole-population, secure data systems supporting SDEs in England



TPP: The Phoenix Partnership (one of the two main primary care computer system suppliers in England); EMIS: Egton Medical Information Systems (the other main primary care computer system supplier in England); IDs: identifiers; NHS#: NHS number; PDS: Personal Demographics Service; MPS: Master Person Service; ONS: Office for National Statistics; IDS: Integrated Data Service.

For details of data sources flowing into secure data systems: see Chapter 3, sections 3.1 and 3.2.

For details of linkage processes in secure data systems: see Chapter 4, section 4.2.

In Wales and Northern Ireland, governments are planning to boost national capabilities with an increasing range of data types and linkages, avoiding the need for separate regional SDEs. The relatively small size of these nations (3.3 and 1.9 million people, respectively) makes this entirely appropriate. A key challenge faced in Wales is how to ensure that the expertise and capabilities of the highly successful SAIL Databank model, which mainly supports the academic research community, can be made more widely available for users from the NHS and – subject to public consultation and relevant approvals – industry. There are currently plans in Wales to develop a new national data platform for access to and use of health and care data for NHS and social care analysis purposes, separate from the SAIL Databank.[273] This may be appropriate if it will provide capability complementary to that of the SAIL Databank, but, given scarce resources, any new data platform capability in Wales must be rigorously justified.

In Northern Ireland, the establishment of a new Health and Social Care Data Institute within Digital Health and Care NI bodes well for further developments of infrastructure for data access and use for health and care research and analysis.[274] However, close partnership and shared data access mechanisms with the Northern Ireland Statistics and Research Agency will be essential if the benefits of linkages between data from health, care and other administrative sources are to be realised.[275] Suspension of the Northern Ireland Assembly over the two years to February 2024 delayed the introduction of legislation to facilitate these developments. Hopefully this will now change. However, clearly articulating the benefits for Northern Ireland's people, health and life sciences sectors, and economy will be needed if any necessary legislative changes are to be implemented in the coming years.

273 See https://dhcw.nhs.wales/national-data-resource/national-data-and-analytics-platform-ndap/.

274 See https://dhcni.hscni.net/digital-strategy/data/.

275 See https://www.nisra.gov.uk/support/research-support/administrative-data-research-northern-ireland-adr-ni.

5.2 Complementary regional secure data environment capabilities

Alongside developments for secure, remote data access at the national level is a growing set of regional SDE capabilities in Scotland and – more recently – England.

In Scotland, a network of four regional data safe havens (SDEs) has existed for several years, enabling access to more granular, often less well-structured data than can be accessed within the national data safe haven.[276] Partnerships between major Scottish universities and NHS Scotland health boards have established these regional safe havens. They vary in their longevity, the range of data available, the size of the population covered, the number and range of research projects supported, the primary location within NHS or university settings, and the resources invested in data storage, curation, access and analysis and user support services. Between them, they cover most – but not all – health boards, and around two thirds of the Scottish population. They are used mainly, but not exclusively, by researchers based within the relevant regions. Given its relatively small size (5.5 million people), Scotland will need to justify ongoing investment in these regional capabilities (and attract additional investment) by seeking to unite them as far as possible into a single national capability, ensuring that they:

- are as meaningfully networked as possible, promoting access from researchers across Scotland, the UK and (where appropriate) internationally;
- seek to extend health board partnerships to cover the entire Scottish population;
- align as far as possible around a single data access process (for example through eDRIS, the electronic data research and innovation service,[277] and Research Data Scotland);
- develop shared metadata and data standards, curation pipelines and analysis support services;
- seek optimal alignment with similar initiatives across the UK.

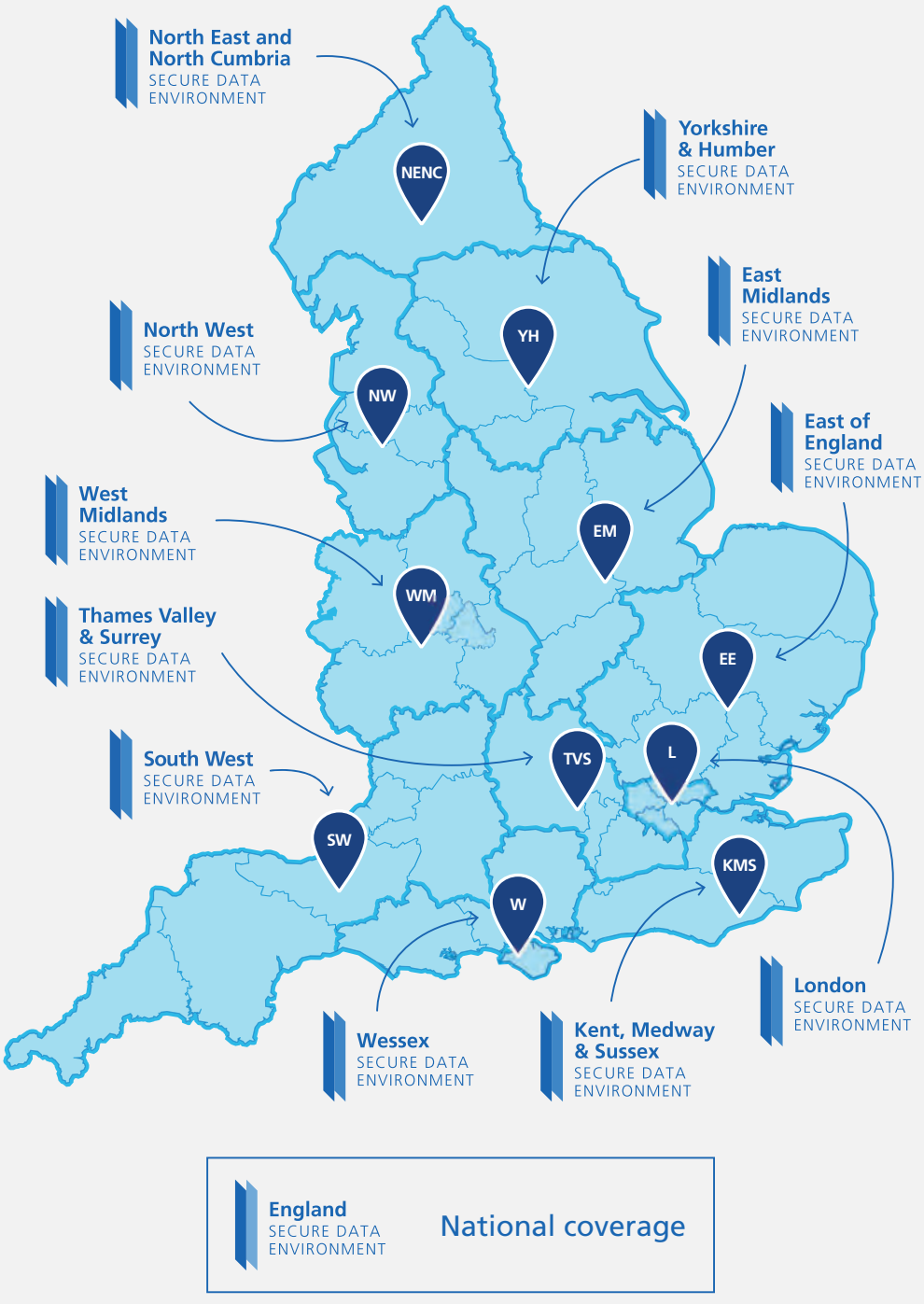
In England, the national NHS England Data for Research and Development programme received HM Treasury approval for a three-year spend of £175 million from 2022/23 to 2024/25, with the goals of further developing the NHS England SDE and NHS DigiTrials service, as well as developing a nationally coordinated network of 11 additional regional SDEs, collectively covering the whole of England (Figure 5.3).[278]

276 See https://www.nhsresearchscotland.org.uk/research-in-scotland/data/safe-havens.

277 See https://publichealthscotland.scot/services/data-research-and-innovation-services/electronic-data-research-and-innovation-service-edris/overview/.

278 See https://transform.england.nhs.uk/key-tools-and-info/data-saves-lives/secure-data-environments/how-will-secure-data-environments-be-delivered/ and https://www.linkedin.com/pulse/transforming-data-enabled-research-landscape-england-bloomfield/.

Figure 5.3 NHS England network of regional secure data environments



The regional SDEs in England are geographically based around clusters of integrated care boards, and together cover the entire geography of England. They have been able to build on previous and ongoing investments, including those in hospital electronic patient record (EPR) systems,[279] regional city-based initiatives,[280] the NIHR-funded Health Informatics Collaborative,[281] and the UK Government Industrial Strategy Challenge Fund's Health Data Research Hubs[282] and digital imaging centres of excellence.[283] These investments have given the Data for Research and Development programme a great head start, and bring considerable expertise, for example in applying natural language processing to extract structured data from medical free text in EPRs, and in curating and analysing radiology and pathology images.

Considerable progress has been made in bringing the SDE network together, with further development of individual SDEs, development of commercial access principles and joint work with Health Data Research UK on a centralised metadata catalogue and 'common front door' to facilitate data access. Remarkable leadership, skill, tenacity and teamwork have been required to make this progress to date. The more recent introduction of joint programme oversight by the NHS England Director of Transformation together with the Government Chief Scientific Adviser for Health (also CEO of the NIHR) signals an important step forward. However, there is a long way to go before the network is fully functional, with substantial challenges to overcome:

1. Some areas of England have already had much more investment and have far more advanced systems than others. Careful balancing of investment and sharing of expertise will be required to avoid exacerbating existing geographic inequalities.

2. These pre-existing developments have occurred largely independently of each other, adding to the challenges of creating a unified network with common approaches to data management, metadata, standards, curation, remote access, analysis support and pricing models.

3. The need for the regional SDEs to complement rather than duplicate the NHS England SDE, adding value through providing access to detailed, granular and unstructured data that are either unavailable – or not suitable to be made available – within the NHS England SDE.

4. The geographical map of the 11 English regional SDEs is primarily based on alignment to England's 42 integrated care boards. Work has already been undertaken to map the regional SDE network to UKRI, NIHR and other major research infrastructure investments (Appendix 7 illustrates the complexity of the task). It also needs to align with England's 29 NHS pathology networks,[284] seven NHS genomic laboratory hubs,[285] and 22 NHS radiology imaging networks.[286] This will need careful planning and engagement to ensure that organisational and geographical complexity is addressed without compromising existing capabilities.

279 E.g. the Local Health and Care Record Exemplar investments: see https://digital.nhs.uk/blog/transformation-blog/2019/so-what-is-a-local-health-and-care-record-anyway.

280 E.g, OneLondon (https://www.onelondon.online/about/) and the Greater Manchester Care Record (https://gmwearebettertogether.com/).

281 See https://hic.nihr.ac.uk/.

282 See https://www.hdruk.ac.uk/helping-with-health-data/health-data-research-hubs/.

283 See https://www.ukri.org/what-we-do/browse-our-areas-of-investment-and-support/data-to-early-diagnosis-and-precision-medicine/.

284 See https://www.england.nhs.uk/pathology-networks/.

285 See https://www.england.nhs.uk/genomics/genomic-laboratory-hubs/.

286 See https://www.england.nhs.uk/transforming-imaging-services-in-england/.

5. The distinction (and any areas of overlap) between the NHS England Research SDE network and its Federated Data Platform (FDP) needs to be clearly laid out to avoid confusion among healthcare professionals, current and potential future data users (such as researchers), patients, members of the public and organisations representing all of these. In brief, our understanding of the FDP is that it is designed for those working for or on behalf of the NHS to deliver and improve patient care. It will provide **operational capability** for healthcare delivery within the NHS in England, while the SDE network will provide data access and services for broader research and analysis purposes. The FDP will focus on improving interoperability across NHS computer systems to enable secure data access to support the care of individual patients. It will also provide data-driven systems to: reduce the backlog of people waiting for appointments or treatments; coordinate care across different parts of the health service; ensure equitable access to vaccination; plan NHS services to meet the needs of the population; and improve efficiency and value for money in NHS purchasing and management of supplies.[287]
6. Some industry-based users with established data access arrangements through partnerships with individual hospital trusts or regional data access platforms that preceded the SDE network expressed concerns to us about the potential for disruption of access while new arrangements emerge. A clear roadmap for transition arrangements as well as for SDE network developments will be essential to manage expectations and minimise any such disruption.
7. As with nationally collated data, mechanisms for the transfer of data out of the regional SDEs to other secure locations, where appropriate and necessary, need to be specified and developed. For example, such mechanisms may be needed to provide linked data for consented clinical studies, longitudinal research cohorts and clinical trials.

5.3 Resources enabling access to general practice data linked to other sources of health data

In England, there are now many platforms providing access to general practice data that cover large amend to: subsets of – or the whole – population and are linked (or linkable) at person level to other data sources. Some of these platforms have been established and have built a substantial user base over a decade or more (for example the Clinical Practice Research Datalink (CPRD)). Others have launched during or since the COVID-19 pandemic, bringing new capabilities (for example OpenSAFELY or NHS England's provision of general practice data within its national SDE – see Box 5.2). These platforms vary with respect to several key characteristics and services, including:

- the host organisation and platform funding arrangements;
- population coverage (whole-country population or a subset);
- mechanisms and capacity to link health-relevant data from multiple sources beyond general practice;
- previous, current – and potential future – scalability (for example the number of researchers or analysts and projects or programmes supported);

287 See https://www.england.nhs.uk/digitaltechnology/digitising-connecting-and-transforming-health-and-care/.

- support for different types of analysis and analyst;
- routes to and speed of data access;
- how close the general practice and other linked data are to being real-time (this is crucial for some but not all analyses);
- extent of support for commercial and/or international data users;
- whether and how data can be securely transferred from the platform to another approved secure location;
- other services provided (for example CPRD's clinical trials services or the Royal College of General Practitioners Research Surveillance Centre's biological sampling services).

There have been major positive developments – and setbacks, discussed extensively elsewhere[288] – in access to and linkage of general practice data in England during and since the COVID-19 pandemic. But, critically, as discussed later in Chapters 6 and 7, none of the current data platforms, either alone or in combination, yet fulfils all the requirements of a national solution for primary care data. Further, a national general practice data solution was raised repeatedly in our discussions with multiple stakeholders across the UK as the highest priority, unfulfilled health data need. And, while those providing and/or contributing to the development of the existing platforms have developed and provide substantial expertise in the management, curation, analysis and uses of linked health data (including primary care data), there is unnecessary overlap and duplication of effort across the system. The current situation is neither affordable nor sustainable and there is considerable potential for more efficient and effective use of existing and future resources.

As regards the devolved administrations, access to general practice data linked to multiple other health-relevant data sources for almost all general practices has been available via the SAIL Databank in Wales for many years. In Scotland access to some general practice data at whole-country scale became possible during the COVID-19 pandemic but has slipped backwards since. There has been progress in developing a single general practice data platform with a range of analysis services in Northern Ireland.

More details of the platforms enabling access to English general practice data that are publicly funded and/or hosted by a public sector organisation are shown in Appendix 8.

288 See https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/general-practice-data-for-planning-and-research/about-the-gpdpr-programme.

5.4 Other publicly funded health data access services

5.4.1 Health Data Research Innovation Gateway

As yet, there is no single, comprehensive, up-to-date catalogue of all sources of health-relevant data across the UK. However, the Health Data Research Gateway is the most comprehensive effort to date. The Gateway has also been designed to coordinate requests for access to data and has the potential to act as a common front door for the UK, steering requestors towards a limited number of streamlined, standardised, data access approvals processes.[289] HDR UK, which runs the Gateway as part of its publicly funded charitable activities, aims to release an updated version with enhanced coverage and search capability during the last quarter of 2024.

5.4.2 NHS DigiTrials

Initially established as a Health Data Research UK Health Data Research Hub,[290] NHS DigiTrials is now a national NHS England data service, being further developed as part of the NHS England Data for Research and Development programme.[291] NHS DigiTrials has already successfully supported several prominent, large-scale research studies, including the RECOVERY and NHS-Galleri trials and Our Future Health (see sections 1.1, 3.3.1, 3.3.4 and 5.4.3). It offers four services to support clinical studies, in particular clinical trials, funded by public, charity or industry sources, that aim to benefit patients and the public. These are:

- Feasibility service: this uses national health data to establish how many suitable people there are in England to take part in a particular trial but does not identify any individuals.
- Recruitment service: this uses national health data together with patient information to identify people who might be suitable for a certain trial, and contacts them to see if they would like to take part.
- Communication service: this provides information on behalf of clinical studies to their volunteers to keep them updated about the progress and results of the study they are participating in.
- Outcomes service: this provides access to data about the trial volunteers from national health databases so that the trial can follow their health over time to demonstrate the short- and long-term impact of trial treatments.

The quality, extent and scalability of the services offered depend on the volume of requests NHS DigiTrials can handle efficiently and effectively, and on the national health data that NHS DigiTrials can use to support its services. Currently, these data are limited to data from hospital episodes, community dispensed medicines, the Personal Demographics Service and the death registry. Some other data sources can be used but only with complex, time-consuming, bespoke work. The incorporation of data from general practices across the country would greatly enhance the feasibility, recruitment and outcomes services.

289 See https://www.healthdatagateway.org/.

290 See https://www.hdruk.ac.uk/helping-with-health-data/health-data-research-hubs/.

291 See https://digital.nhs.uk/services/nhs-digitrials.

5.4.3 Longitudinal research resources

Some very large and widely used population-based research resources in the UK, UK Biobank, Our Future Health and resources held by Genomics England, have made major investments in obtaining linked health-relevant data from a range of different national sources. They have also invested in enabling researcher access to the data they hold, together with the linked health-relevant data, via their own highly specialised SDEs.[292, 293, 294]

However, there are many other population-based longitudinal studies across the UK, collectively including around two million research participants. The effort, time and cost required to request, obtain access to and link health-relevant data from the participants in each of these studies, as well as to provide secure access to the cohort and linked health data for a wide range of researchers conducting research for public benefit, is considerable.

Established in response to this challenge, the UK Longitudinal Linkage Collaboration (UK LLC) is a collaborative endeavour involving many of the UK's most established longitudinal studies.[295] The UK LLC is led by the Universities of Bristol and Edinburgh, in collaboration with University College London, Swansea University's Secure Research Platform UK, and the University of Leicester. It provides a national secure data environment for longitudinal population-based research together with a data linkage service and resource to its partner studies and a simple one-application process to UK-based researchers applying to access linked longitudinal data. As a result, it enables cross-sector research and supports researchers to respond to immediate and future policy needs. The greater availability of large-scale, diverse linked data from across multiple cohorts will help researchers to study rarer outcomes and seldom reached populations.

A similar initiative has been established by the British Heart Foundation Data Science Centre at Health Data Research UK. The BHF Data Science Centre cohorts platform aims to provide a data linkage service for the very large number of health condition focused longitudinal cohort studies across the UK, together with a secure data environment for safe, remote researcher access.[296]

292 See https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform.

293 See https://ourfuturehealth.org.uk/protecting-your-data/how-we-make-data-available-for-research/.

294 See https://www.genomicsengland.co.uk/research/research-environment.

295 See https://ukllc.ac.uk/.

296 See https://bhfdatasciencecentre.org/areas/cohorts/.

5.5 Secure data environment accreditation and standards

5.5.1 The Five Safes Framework

For research and analysis, the widely used and internationally accepted 'Five Safes Framework' (**safe data, safe research, safe people, safe settings, safe outputs**) was designed by UK experts to protect the privacy and security of people's data, to ensure that data are used for the public good, and to guard against misuse.[297] Most SDEs operate under the principles of the Five Safes framework. Where possible, data custodian organisations either de-identify or irreversibly anonymise data (**safe data**) before making them available for approved uses for public benefit (**safe research**) to appropriately trained, certified and authenticated analysts (**safe people**) within SDEs (**safe settings**). Before analysis results (for example tables or figures) are exported, they are checked to ensure that they could not be used to identify any individual (**safe outputs**).

5.5.2 Accreditation of SDEs

The Digital Economy Act 2017 (DEA) facilitates the linking and sharing of de-identified data by public authorities for accredited research to generate new insights about UK society and the economy. The UK Statistics Authority (UKSA) is the statutory body responsible for the accreditation of processors, researchers and their projects, following the principles of the Five Safes Framework.[298] The UKSA uses robust criteria to ensure that accredited processors (i.e. secure data environment providers) meet cross-government standards for securely holding sensitive data (personal information); use appropriate technical infrastructure; publish and maintain appropriate data policies; have appropriate skills and experience; and are listed on a public register.[299] DEA-accredited processors must be able to de-identify data before making them available to accredited researchers in a secure environment and must ensure that any analysis results exported by researchers from the environment are 'disclosure controlled' so that that no individual represented in the data could be identified.

Although the DEA excludes sharing of NHS-held health and social care data, the UKSA has accredited several UK SDEs that hold health and care data as well as data from other sectors, for example the Welsh Secure Anonymised Information Linkage (SAIL) Databank and Scotland's electronic Data Research and Innovation Service (eDRIS), which works in partnership with the Edinburgh Parallel Computing Centre (EPCC) to host Scotland's national data safe haven.

In 2015 the Scottish Government developed a charter that set out the agreed principles and standards for the routine operation of its federated network of one national and four regional safe havens (that is, SDEs) in Scotland where data from electronic NHS patient records are processed, linked with other data and analysed to support research, while protecting patient identity and privacy, when it is impractical to obtain individual patient consent.[300] The Scottish Government has now commissioned Research Data Scotland and the Scottish Save Haven Network to update this charter. This will take into account advances

297 See https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/ and https://digital.nhs.uk/services/secure-data-environment-service/introduction/five-safes-framework.

298 See https://uksa.statisticsauthority.gov.uk/digitaleconomyact-research-statistics/.

299 See https://uksa.statisticsauthority.gov.uk/digitaleconomyact-research-statistics/better-access-to-data-for-research-information-for-processors/list-of-digital-economy-act-accredited-processing-environments/.

300 See https://www.gov.scot/publications/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/pages/4/.

in technology (for example the need for SDEs to be able to support increasingly advanced analysis methods including AI) and the need for more streamlined processes to support the rapid generation of policy-relevant insights.[301]

The establishment of the NHS England SDE, OpenSAFELY, and the associated network of regional NHS SDEs across England is more recent (see sections 5.1 and 5.2). Although these operate within the Five Safes Framework, none has been accredited by the UKSA. DHSC has committed to establishing a robust accreditation regime for SDEs and is working with the UK Statistics Authority (UKSA) to put in place a new, amended version of its accreditation framework by spring 2025.[302] This is encouraging. However, a UK-wide SDE accreditation system for SDEs, whether or not they hold health and care data, will be crucial to enable cross-sectoral data access arrangements that extend across the UK. This will require the involvement of the devolved administrations in the design and implementation of the proposed accreditation framework.

5.5.3 Technical standards for SDEs

With more SDEs being set up all the time, a UK-wide system for standards as well as accreditation of SDEs will accelerate the safe use of health data for patient and public benefit. The Standard Architecture for Trusted Research Environments (SATRE) project,[303] supported by UK Research and Innovation's Data and Research Analytics Environments UK (DARE UK),[304] aims to provide such a set of UK-wide standards. It incorporates knowledge and best practices from multiple institutions and sectors across the UK. It covers all aspects of SDE provision, including information governance procedures, computing technology, data management and other supporting capabilities. It aims to standardise the capabilities of SDEs, making it easier for users, operators and developers to work with sensitive data, and making the operation of SDEs more transparent to data owners and the general public. SATRE's guidance is based on four key principles for SDEs:

- Usability: SDEs must minimise barriers to use, balancing the trade-off between operational security and usability to provide a productive and accessible analysis environment.
- Maintaining public trust: SDEs should build and maintain the trust of data subjects and any other impacted individuals, groups, communities and organisations by protecting privacy, keeping data secure and being transparent about their work.
- Observability: Human initiated and automated processes within an SDE should be observable and auditable to ensure that policies and controls are doing what is intended.
- Standardisation: SDEs should adhere to standards wherever possible, making it easier to design, operate, use, understand and audit SDEs, and reducing duplication of work.

301 See https://www.researchdata.scot/news-and-insights/coming-soon-the-scottish-safe-haven-charter-20/.

302 See https://www.gov.uk/government/publications/data-access-policy-update/data-access-policy-update.

303 See https://satre-specification.readthedocs.io/en/v1.0.0/.

304 See https://dareuk.org.uk/.

Chapter 6

Priorities, barriers and solutions

In this chapter

National health and data bodies should focus on several key priorities to maximise the benefits of using heath data. Our focus in this chapter is on England. However, most priorities and barriers, and many solutions, are common across all four nations of the UK. Our extensive discussions yielded remarkable consistency across the wide range of stakeholders we consulted with on their priorities for:

- the changes or improvements needed in the UK's health data ecosystem to facilitate timely, secure access to linkable sources of health-relevant data for public benefit;
- the types and sources of data that are either inaccessible or not as accessible as they should be to maximise patient and public benefit.

6.1 System priorities

Policy, research and health service needs during the COVID-19 pandemic drove major gains in efficiency, productivity and positive impact on patient care and public health policy of improving the flows, linkage and secure accessibility of linkable health data through a limited number of national data custodians (noting that each additional data custodian adds complexity, expense, delay and increased potential for error). But our broad-ranging consultation confirmed that many of these gains have been temporary. Indeed, in some respects we are slipping backwards, and critical system gaps remain. These system gaps make some crucial national tasks difficult and at times impossible. Some examples are shown in Box 6.1.

Box 6.1 Examples of crucial national tasks that are difficult or impossible due to health data system gaps

National drug and device health safety monitoring

After a new drug or device has received regulatory approval for use in clinical practice, the UK's medicines and medical devices regulator, the Medicines and Healthcare Products Agency (MHRA), needs mechanisms for the ongoing monitoring of drug and device safety. These are needed for the rapid detection of expected and unexpected adverse effects of medicines and medical devices. This in turn informs guidance on the use of medicines and medical devices in different types of people according to age, sex, medical history and other characteristics. In some situations, nationwide monitoring across the whole population is important (for example for monitoring potential rare adverse effects of widely used vaccines). However, the MHRA cannot currently access all the sources of national data needed for the most appropriate monitoring of the safety of all drugs and devices. It can rapidly and efficiently access and analyse high-quality data from a subset (around 30%) of general practices via the Clinical Practice Research Datalink, a resource that it hosts (see section 5.3 and Appendix 8). But it cannot readily access, when needed, national-scale data:

- from all – rather than only some – general practices (see sections 3.1.2, 5.3 and Appendix 8);
- with real-time information from hospitals on diagnoses, procedures and laboratory test results (see section 3.1.4);
- on drugs prescribed in hospital and high-cost drugs (see section 3.1.5);
- on the unique identifiers of all the many different implantable medical devices (for example artificial replacement heart valves or joints) used in healthcare.

These data access issues make it difficult for the MHRA and others to rapidly detect, track and further interrogate serious adverse events that may be associated with new drugs or devices. Similar problems hinder the work of other organisations that need to access and analyse data to better understand drug and device safety, including national and regional NHS organisations, the National Institute for Health and Care Excellence, drug and device manufacturers, and researchers working to inform all of these. These issues must be addressed if the recommendations made by Baroness Cumberlege in her 2020 '*Independent Medicines and Medical Devices Safety Review: First do no harm*'[305] are to be fully implemented.

305 See https://www.gov.uk/government/publications/independent-medicines-and-medical-devices-safety-review-report.

Tracking and responding to epidemics and pandemics

During the COVID-19 pandemic, the importance of accurate and up-to-date monitoring of infectious diseases and vaccine uptake across the whole population of the UK became widely known. The difficulties in accessing the different sources of data needed to do this are less widely appreciated.

For example, as part of its vital role to maintain our safety and security by monitoring infections across the population, the UK Health Security Agency (UKHSA) needs to analyse microbiology laboratory data on infection test results from all relevant testing laboratories across the country (see section 3.1.6). This is to find out which people have had a test and which have tested positive for the infectious diseases being monitored. The reporting system works well for some purposes and some infections. However, these laboratories are mandated to provide data to the UKHSA about some – but not all – infectious diseases. The result is that the data for comprehensive monitoring of some infections will be incomplete. In addition, these laboratories are mandated to provide information on positive but not negative test results, which means that the UKHSA cannot always reliably track the proportion of people having a test who test positive, an important measure for infectious disease monitoring.

It is also important for the UKHSA to be able to monitor how characteristics such as age, ethnicity, geographic location and deprivation affect not only the risk of acquiring an infection but also of the uptake of vaccines that are part of national vaccination programmes. To do this effectively, the UKHSA needs to be able to access data rapidly and at whole-population scale from several different sources. One important source is data from all general practices (see sections 3.1.2 and 5.3). These were available to the UKHSA – via NHS England's General Practice Data for Pandemic Planning and Research dataset – for COVID-19-related vaccine uptake analyses. However, they are not available to inform the monitoring of uptake of other important vaccines such as those against measles, whooping cough and a range of other infectious diseases.

Stakeholders' highest priorities for improvements to address these system gaps in the health data ecosystem are detailed further in Appendix 9. In summary, there is a need to:

- **Increase speed, timeliness and scope of data access.** Data access processes remain unacceptably slow and tortuous (see Box 6.2 for some illustrative examples). There are many reasons, which need to be addressed through a range of solutions to drive a significant improvement on the current situation. Improving the capacity and processes for provision of linked health data from NHS England and other major health data custodians would help to reduce the delays and their associated costs, as well as bringing the benefits of health research to patients and the public more rapidly.
- **Maintain broader access achieved during the pandemic.** The UK's response to the pandemic would have been better informed if better data flows, linkage and access had been in place beforehand. Co-operation and collaboration between multiple national organisations, aligned around a common goal, together with a more proportionate approach to balancing the benefits versus the risks of data use, brought rapid improvements in data access and linkage during the pandemic. These enabled insights from data that rapidly informed patient care and health policy. However, there is an increasing risk of losing these advances through drift back to pre-pandemic ways, loss of clear alignment of incentives, the effects of NHS organisational change, wider political upheaval and fiscal challenges.
- **Maintain and enhance national data assets.** There are strong arguments for prioritising access to key national datasets from different sources relevant to a wide range of health conditions (notably general practice, hospital, medicines, and mortality data). These should be a foundation onto which more specialist, domain-specific national data can be layered (see section 6.2). Given limited resources, it will be important to ensure that more recent investments in secure systems for access to regional (as opposed to national) data across England avoid duplicating these national data efforts and instead focus on enabling access to more granular, unstructured data that cannot yet be brought together readily at whole-country scale, especially in England.
- **Maintain and improve capability for secure data transfer**. All four nations of the UK are moving towards a position of data access within secure data environments (SDEs) as the default access route, to minimise unnecessary movement of data and to enhance security and privacy. While this makes a great deal of sense, the capability for secure transfer of data to secure locations outside of NHS environments must be preserved. This will allow the transfer of linked data, usually underpinned by explicit participant consent, to the secure settings used or hosted by the UK's internationally recognised research cohorts and clinical trials.[306] Many of these are hosted by large research organisations, working closely with but holding data in secure environments external to the NHS. It will also facilitate the sharing of health data with accredited non-NHS SDEs, for example the ONS Integrated Data Service (or its forerunner the Secure Research Service), enabling linkages with health-relevant data from non-NHS settings.

306 There are literally hundreds of these, but good examples are UK Biobank (https://www.ukbiobank.ac.uk/), Our Future Health (https://ourfuturehealth.org.uk/), Genomics England's 100,000 genomes project. (https://www.genomicsengland.co.uk/initiatives/100000-genomes-project), the Recovery trial (https://www.recoverytrial.net/) and the NHS-Galleri trial (https://grail.com/clinical-studies/nhs-galleri-trial-clinical/).

- **Improve data usability.** Better quality data and metadata,[307] improved interoperability of health, care and other computer systems (see Chapter 3, especially section 3.1.1[308]), and the standardisation of data collection, formats, and terminologies between and within UK countries will all help to enhance the efficiency, accuracy, reproducibility and relevance of analyses and insights. As highlighted in the Goldacre Review (see Appendix 3), the implementation of open, shareable and reproducible approaches to data management, curation and analysis pipelines will drive transparency and efficiency, and reduce duplication of effort. For example, the Clinical Practice Research Datalink (see section 5.3) has developed expertise in curating data from general practices and other sources for use by the MHRA and a large numbers of external research users based in academia and industry. Several other examples of more recent innovative advances exist. These include those provided by OpenSAFELY and the BHF Data Science Centre (see section 5.1) in their work with NHS England to enable efficient, secure access and use of multiple sources of linked health data at whole-population scale. Some regional SDEs have also developed exemplary data management, curation and analysis pipelines.
- **Make SDEs the most attractive option for most uses and users.** We endorse the move towards health data access and analysis within SDEs, wherever possible and appropriate (noting the need for secure data transfer to other secure locations covered earlier). However, these environments – and the data held within them – must be user-friendly with scalable user support services, allow a range of analytic approaches, including machine learning and AI, and have transparent and affordable costs. This will ensure that users embrace the change with enthusiasm and gain rather than lose momentum in analysis productivity.
- **Improve transparency for and meaningful engagement with patients, public and healthcare professionals, policymakers and politicians.** It cannot be emphasised enough that involving and engaging patients and members of the public in discussing and overseeing the benefits, risks and governance of access to health data is essential. Transparency with policymakers and politicians is also crucial. This is not a one-off process. It must be an ongoing interaction, as the landscape is constantly evolving. None of these groups can engage meaningfully without clear information. Clarity is particularly difficult when the landscape is so complex. Reducing complexity is one of the many ways in which transparency – and so meaningful engagement and involvement – can be improved.

For each of these system priorities, Appendix 9 lays out what is needed and why, the main barriers and potential ways in which these might be overcome.

307 Metadata is information about datasets and the data items within them, including data dictionaries and descriptions of other characteristics such as coverage and missingness.

308 Mandatory compliance with specified information standards by suppliers of IT systems used for processing health and/or adult social care information may be included in UK Government's Data (Use and Access) Bill (https://bills.parliament.uk/bills/3825). This could help with interoperability challenges.

Box 6.2 Costly delays in access to health data for research and analysis

Total annual expenditure on health research and development in the UK is around £10 billion. Approximately half is accounted for by public or charitable expenditure.[309] A significant and increasing proportion of this funds clinical trials and other research studies that depend on access to health-relevant data generated within and beyond the NHS.

We heard from research funders and researchers of many examples of research studies that were held up or abandoned because of delays in access to health data. Resolving these delays could generate considerable cost savings, particularly important for major national research funding bodies and charities that draw on scarce public funds (via taxes) and donations.

Two-and-a-half years to link data from five national datasets for research to improve services for congenital heart disease patients

One group of researchers described a two-and-a-half-year process to link data from five national datasets for their research study to improve services for patients with congenital heart disease. They concluded that "NHS data can inform and improve health services and we believe there is an ethical responsibility to use it to do so," but that "The current system is incredibly complex, arduous and slow, stifling innovation and delaying scientific progress."[310]

Delays of months to years to obtain linked health data for multiple research studies funded by the National Institute for Health and Care Research (NIHR)

NIHR is the largest single funder of health research in the UK. Its annual research spend is £1.3 billion. Almost 10% of this is spent on its Health and Social Care Delivery Research and Health Technology Assessment programmes. These fund research to improve the quality, accessibility and organisation of health and social care services, and to generate evidence on the clinical effectiveness, cost-effectiveness and broader impact of treatments, tests, and other interventions in health and social care.

NIHR colleagues told us about 29 studies funded by these two programmes that were delayed during the 2022/2023 financial year because of waits of several months or more to obtain health data on the studies' participants from NHS England. These studies are testing medical, surgical and other interventions for cancers, kidney disease, liver disease, heart disease, stroke, infectious diseases and frailty in older people; investigating delivery of care for babies, children and young people; and assessing health service use among people from ethnic minority groups. The delays have resulted in administratively costly extensions to funding awards of up to two years. They have also delayed the emergence of research findings that could save and improve many people's lives through changing healthcare practice and public health policy.

309 See UK Clinical Research Collaboration (2023). *UK Health Research Analysis Report 2022* (https://hrcsonline.net/wp-content/uploads/2024/04/UK_Health_Research_Analysis_Report_2022_web_v1-1-postpub.pdf).

310 E.g. see Taylor JA et al. *The road to hell is paved with good intentions: the experience of applying for national data for linkage and suggestions for improvement*. BMJ Open 2020 (https://bmjopen.bmj.com/content/11/8/e047575).

Similar delays are likely to affect many other parts of the NIHR funding portfolio as well as many research studies funded from other sources. The delays reported here relate to provision of data from NHS England, but delays in obtaining data from other health data custodians across the UK will also affect these and other health research studies.

Waiting for over a year to link existing national datasets to study the impact of environmental exposures on children's health and educational outcomes

The publicly funded Kids' Environment and Health Cohort study aims to link vital statistics, health, education and census data for all children born in England since 2006 with data about the local environment in and around children's homes and schools. Once linked, these data will enable multiple research studies to inform policy on issues such as:

- the impact on children's health of in-utero exposure to air pollution or of living in poorly heated homes;
- the mental health and educational consequences of attending schools near gambling outlets;
- the effects of low emission zones on respiratory infections and antibiotic use in children.

Funding for the initiative started in December 2022. Shortly afterwards, the research team agreed an approach to data linkage with NHS Digital and the ONS and sought relevant approvals. Over a year later, in March 2024, NHS England advised that the agreed linkage approach would not be possible after all. An alternative approach was agreed and the research team sought a revised set of approvals. However, NHS England informed the team in September 2024 that their application had been paused due to its complexity and NHS England resource constraints.

These delays are preventing essential, publicly funded research to better understand the wider determinants of health and wellbeing in children.

6.2 Data priorities

Stakeholders' main data priorities are summarised in Appendix 10. These focus on:

- sources of data that are already collected and collated at **national scale**, but not yet accessible in the ways that would bring most benefit;
- national data assets that do not yet exist but that are perceived to be achievable and desirable.

In Chapter 3, we described the wide range of datasets arising from health service and broader administrative activities that are already collected nationally from various sources and collated by national data custodian organisations operating across the four nations of the UK.

These nationally collated datasets include both **'generic' datasets,** relevant to research across multiple disease domains, and **'domain-specific' datasets**, providing more detailed information on specific health conditions. Generic examples include data from national death registries (providing information on cause and date of death), hospitals (providing data on diagnoses and procedures occurring during hospital admissions), coded general practice data and data on medicines dispensed from all community pharmacies. Domain-specific examples include data from the UK's well-established national cancer, cardiovascular, diabetes, renal, respiratory and joint replacement audits and registries.

It is of note that these national datasets are mainly highly **structured** (see Box 3.1). Structured data represent less than 20% of the data arising from NHS healthcare, most of which are unstructured and complex (for example data from free text medical correspondence or radiology images). But the breadth, depth and scale of coverage of the structured national data (i.e., including information about all health conditions affecting people across all age groups, ethnicities, social backgrounds and geographic locations), mean that they are of extraordinarily high value for addressing a wide range of questions and generating diverse insights. Further, in comparison with unstructured data, they are more straightforward to store, de-identify or anonymise, transfer, link and access securely. As a result, most people and organisations we consulted see them as the lowest hanging fruit of our national health data assets.

Hence, a major priority is to ensure streamlined access to key generic national structured datasets from general practice, hospitals, death registries and medicines data sources, linked to each other at population-wide scale, providing a foundation onto which additional, more domain-specific and more complex, unstructured data (such as data from NHS images) can be layered. Almost everyone we consulted highlighted as their top priority fulfilling the ongoing need for access to and linkage of comprehensive, coded, structured general practice data from across the whole populations of each of the four nations of the UK.

However, the 80% of healthcare data that are **unstructured** represent a significant, under-utilised resource. While most stakeholders we consulted prioritise access to structured data, where achieving national scale is technically less challenging, many also want to see advances in access at scale to unstructured health data. Although such data are more complex, they add substantial detail and granularity, as well as greater diversity of data types. Despite the challenges, there has been progress in generating accessible, linked, curated collections of imaging data from routine NHS activity with increasing diversity and scale, and accruing evidence of their impacts (see section 3.1.7).

An ambition to further scale existing NHS imaging data access and analysis capability is one that several individual and organisational stakeholders raised as a priority. Access to other types of unstructured data, for example the information buried in free text of medical notes, correspondence, and specialist reports on radiology and pathology tests, is also of interest to many stakeholders. This type of information is not available at national scale, but can increasingly be accessed, alongside structured data, within regional, mainly hospital-based, SDEs (see section 5.2). Pharmaceutical, data and digital technology and commercial clinical trials companies often need access to more granular, regional data. For example, this may be necessary to identify patients for recruitment to clinical trials in cases where national data are insufficiently detailed, or to develop and evaluate innovative tools and technology solutions (such as natural language processing tools that automatically extract structured data from medical free text, or AI systems to streamline complex radiology workflows) as close as possible to the real world healthcare front line where they are intended to be used.

Beyond the NHS, administrative data from other sectors (for example social care, education, census, disability, income, and justice data) have enormous yet still largely untapped potential. We have seen that data from many of these sources are already collected and collated at national level by relevant government departments (section 3.2). Insights of huge relevance to the public's health will come from linkage of these health-relevant data to data from health and care settings. Analyses of such linked data would improve understanding and inform policies about the broader determinants of health and wellbeing, health inequalities, and the consequences of poor physical and mental health. However, except for the SAIL Databank in Wales, progress has to date been largely restricted to a series of bespoke linkage projects, rather than the implementation of a systematic, routine approach. This current situation limits the efficiency and scale of beneficial outputs, particularly in England.

A summary of these data priorities (detailed further in Appendix 10) is as follows:

- **General practice data:** comprehensive coded general practice data at national scale in near real time, accessible and linkable for the full range of beneficial uses.
- **Hospital emergency department and admissions data:** enhance existing national hospital episodes data with more granular diagnostic coding and access in near real time.
- **Medicines data:** comprehensive national-level data on medicines prescribed and dispensed in hospital, and on high-cost medicines, in near real time – to complement data now available on community prescribed dispensed medicines.
- **Hospital outpatient data:** mandatory inclusion of diagnostic and procedural codes in national hospital outpatient episodes data, supporting analyses and insights across all health conditions, whether managed in hospital or not.
- **Laboratory data:** comprehensive national system(s) for data on laboratory assay test requests and results, extending established national microbiology and NHS genetic sequencing national laboratory data capability.
- **National audits and registries data:** domain-specific, national (but often siloed) data, accessible and linkable via centralised, coordinated national data custodian processes, to enhance research studies, increase data quality through wider use and reduce duplication of data collection efforts.

- **Screening data:** national screening programme data accessible via centralised coordinated national data custodian processes, linkable to other health data, including on health outcomes, enabling the evaluation of impact of screening on health outcomes and of more targeted inclusion criteria for screening.
- **Social care data:** national adult and children's social care data accessible and linkable via centralised, coordinated national data custodian processes, enabling analyses to understand demand for, equity of access and provision of social care, as well as the health determinants and outcomes of different types of social care provision.
- **Other cross-sectoral data:** streamlined, scalable processes for national linkages of NHS data to health-relevant data from across other government sectors, allowing analyses of the wider determinants and consequences of health and healthcare, with policy relevant insights that benefit patients and the public.
- **Imaging data**: large-scale population-based imaging resources based on routine NHS imaging, securely accessible and linked to other health data for development and testing of automated imaging processing and analysis tools (many AI-based), and better characterisation of participants in research studies.
- **Other unstructured data**: improved access to unstructured data (such as free text) alongside structured data, bringing substantial added granularity and proximity to the clinical coalface.

For each of these data priorities Appendix 10 lays out what is needed and why, the main barriers, and potential ways in which these might be overcome.

6.3 Summary of key barriers and potential solutions

For the key stakeholder system and data priorities outlined in sections 6.1. and 6.2, Appendices 8 and 9 lay out some of the main barriers and potential solutions for overcoming them. Here we summarise these, drawing on other sections of this review where relevant. While the focus here is on England, we should emphasise that most of the barriers, as well as the principles behind the suggested solutions, are common across all four nations of the UK.

6.3.1 Addressing system priorities

It would be easy to suggest that difficulties and delays with access to and linkage of health-relevant data relate mainly to chronic under-investment, and that increased investment is the main solution. However, while additional strategic investment in particular areas is part of the solution, significant progress will only be made through recognising other key barriers, and addressing these with a combination of financial, political and technical solutions.

A major challenge is the present ecosystem complexity and fragmentation, which applies within and across the NHS in all four nations (see section 3.1.1) but also to the many non-NHS organisations involved in the health data landscape. Alignment of national organisations around common goals and priorities was behind the delivery of several health data-driven initiatives during the COVID-19 pandemic that benefited people globally, some occurring at unprecedented pace and scale. Similar approaches must now be adopted and enhanced to tackle the UK's other epidemics, as well as global pandemics of obesity, cancer, diabetes, dementia, cardiovascular disease, mental health conditions and others. The relevant national organisations must commit to:

- a coordinated joint strategy that recognises health data as a critical national infrastructure, since the UK's health data ecosystem and the intelligence it does or should generate is essential not only to people's health but also to their safety, security and economic stability;[311]
- reducing unnecessary complexity and duplication of effort (and spend);
- supporting a cultural shift towards a health data ecosystem that promotes the uses of health data for patient and public benefit, that rewards and incentivises behaviours that drive this, and that holds senior NHS England, National Institute for Health and Care Research (NIHR) and UK Research and Innovation (UKRI) leadership accountable for enabling health data-driven research and analysis;
- backing the appointment of a highly credible senior executive leader at the highest level within NHS England, NIHR/DHSC and UKRI, with responsibility and ring-fenced budget to deliver a national health data service to support health data-enabled research and analysis.

In the current economic climate, expectations around investment must be realistic. Many opportunities exist for savings, for example through reducing complexity and duplication, and by standardising and streamlining data access via a single national health data access system. These would reduce existing costs and allow better use of resources, such as those currently needed for large teams of staff to navigate overly complex data access processes, or to administer extensions to research funding awards required because of lengthy delays in access to data (for examples see section 3.3.2). As important is the need for a change in the culture of data custodian and controller organisations, which should be driven through incentives and performance targets that reward the facilitation of rapid secure access to data and the provision of services to improve the productivity of data users.

Other significant challenges are created by the lack of long-term funding for national health data infrastructure initiatives. Where funding has been made available, for example through the recent NHS Data for Research and Development programme, timeframes for delivery have been tight and exacerbated by delays in the release of funds. Delivery has also been hampered by significant organisational turbulence and capacity issues within NHS England. The recent merger of NHS England, NHSX, NHS Digital and Health Education England has diverted focus for many months to organisational restructuring, and has led to substantial reductions in staff numbers, including in information governance and specialist data management and curation teams. Ongoing headcount caps and skills shortages restrict the ability to recruit essential new staff. NHS England can only deliver the much-needed national health data infrastructure through strategic partnerships with external organisations. A solution is joint accountability (including for a specific set of performance metrics), with NIHR and UKRI as major public health research funding bodies, of a national service to support health data-enabled research and analysis. The UK's national institute for health data research, Health Data Research UK, which has reach across the UK's universities, is already playing a major role in supporting the delivery of national data infrastructure through informal partnerships with NHS England. Formalising these partnerships would further help to bring the necessary expertise and connectivity to research users. It would also provide a route

311 From a UK government perspective, critical national infrastructure means those facilities, systems, sites, information, people, networks and processes necessary for a country to function and upon which daily life depends. The disruption of such infrastructure would impact public safety, security, health or economic stability. See https://www.npsa.gov.uk/critical-national-infrastructure-0.

to filling the substantial capacity gaps, for example through secondments into NHS England. Involving organisations representing potential future commercial users of such a service, for example the Association of British Pharmaceutical Industry, in the design and delivery of this national health data service would bring not only additional expertise but also the potential for precompetitive industry investment that would benefit all users and enhance patient and public benefit.

Many people and organisations we spoke to in consulting for this review raised concerns about the potential for current and future NHS Data for Research and Development programme investment in regional data infrastructures (specifically the regional SDEs) to divert resources and focus from national infrastructure and data capabilities, which are crucial for some of the UK's most successful life sciences initiatives.[312] However, like researchers in universities, regional NHS organisations (such as hospital trusts or regional groupings of these) have encountered delays or barriers in gaining access to national data (or relevant regional slices of national data). These regional organisations will not support investment in national infrastructure that does not fulfil regional needs. Regional health data infrastructures should benefit from what exists (or can readily be created) nationally, allowing them to focus on bringing added value through the depth and granularity of data that cannot be provided through national data collections.

A further frequently raised concern was that the policy move to access to data within SDEs as default might mitigate against the maintenance and enhancement of mechanisms for the secure transfer of data out of NHS settings for specific legitimate purposes, recognising that the NHS Research SDE Network will never be able to support all data analysis needs. It is encouraging that this requirement has been recognised in the Department of Health and Social Care's data access policy.[313] However, efficient and secure mechanisms for data transfer from regional as well as national secure environments need to be developed, enhanced and maintained. The increasingly international nature of both research and the life sciences sector means that consideration and robust solutions for data access and transfer beyond the UK are also needed.

The proposed national health data service will also need to address several additional technical challenges, as laid out in Appendix 9. These include improved data usability,[314] improved SDE user experience that make SDEs a desirable option for data users (including, for example, the capacity to support advanced analyses using AI), and contributing to the development of SDE standards and accreditation schemes that are accepted and implemented UK-wide (see also sections 5.5.2 and 5.5.3).

312 E.g. the UK's internationally leading genomic resources, population-based cohorts, clinical studies and clinical trials.

313 See https://www.gov.uk/government/publications/data-access-policy-update/data-access-policy-update and https://www.gov.uk/government/publications/secure-data-environment-policy-guidelines.

314 E.g., through improved metadata, standardised data formats and terminologies, and data linkage at both person and household (Unified Property Reference Number) level, with record level information on match quality.

As noted in section 6.1 and Appendix 9, building and maintaining trust with patients, public and health professionals will depend on meaningful, ongoing engagement, together with the provision of clear, consistent and accessible information. This needs to promote awareness of the benefits of a wide range of uses of health data for patient and public health, wellbeing and economic productivity, as well as addressing potential concerns and questions about privacy, security and choice. Transparency and consistency of information can only be helped by reducing complexity, especially of data governance and access processes. Engaging with health professionals, particularly GPs, is crucial, given that some have reservations about access to and uses of the health data they collect (see section 2.3). Developing a single, centralised and accessible system that allows people in England to opt out (and to opt back in) of sharing of their data for different types of purposes beyond the delivery of clinical care is a further critical need. This must not impose a burden on GPs and should abandon the currently confusing and cumbersome separate opt-out systems for different sources of health data (so-called type 1 and type 2 or national opt-outs).[315]

From a legal standpoint, more consistent interpretation and application of existing legislation would help to reduce the complexity of the processes to access, link and analyse health-relevant data. There are also some potential legislative changes that could help to simplify these processes. These include:

- Updating the Control of Patient Information (COPI) Regulations 2002. These regulations do not fully reflect how the use of data has developed since they were established over 20 years ago. Updating them could simplify and streamline the processes for enabling health and care organisations to use and share patient information for healthcare delivery and for wider research and analysis purposes. This would build on the successful use of COPI notices during the COVID-19 pandemic to support uses of data for patient and public benefit.
- Amending the Digital Economy Act 2017 (DEA). Including health and care bodies in the provisions of the DEA could reduce the barriers to some types of linkage and analysis of cross-sectoral health-relevant data for public benefit. However, any such amendments would require careful consideration, as discussed further in the following section (6.3.2).

6.3.2 Addressing data priorities

Generic data – essential information on a wide spectrum of health conditions

As discussed in section 6.2, the greatest need is to ensure access to data that provide information across the spectrum of health conditions managed in the community as well as in hospitals. The most pressing need, consistently raised by almost all stakeholders, is for a national solution for access to comprehensive, coded general practice data that can be linked at whole-population level to other sources of health data and used for a wide range of purposes for patient and public benefit. Hospital episodes data need to be enhanced in granularity and timeliness. And solutions should be developed for access at national scale to hospital-prescribed and high-cost medicines, to laboratory data requests and results (including genomic data), and to radiology and pathology imaging data.

Of these 'generic' data priorities, structured, coded general practice data are straightforward technically because, for England, these are almost all held within the systems of just two

315 See https://www.nhs.uk/using-the-nhs/about-the-nhs/opt-out-of-sharing-your-health-records/ and https://understandingpatientdata.org.uk/your-choices.

commercial computer system suppliers and, although rich in content, are low in data volume (see section 3.1.2 and Box 3.2). A key barrier to access relates to the distribution of responsibility for these data across each of over 6,000 English general practices, which take their associated professional and legal responsibilities (and the associated liabilities) very seriously. Secretary of State directions under section 254 of the Health and Social Care Act 2012, section 251 of the National Health Service Act 2006 and notices under the Control of Patient Information Regulations 2002 have facilitated access to and linkage of general practice data for COVID-related healthcare planning and research purposes. These have provided security and confidence to – as well as placing a legal requirement on – general practices to permit data sharing to support the response to the pandemic. This has resulted in many COVID-related uses of general practice data that have benefited millions of people, with no breaches of privacy or security. Extension of these mechanisms to cover non-COVID related analysis and research, so supporting the response to the pandemics of obesity, diabetes, cancer, cardiovascular disease, dementia and others, would enable a far wider range of uses of general practice data than is currently possible. In Chapter 7, we propose a rapidly implemented interim solution to fulfil this pressing need while options for a definitive national general practice data solution are worked through. Success will rely on engaging GPs throughout, avoiding imposing any burden on already overstretched primary care services, and providing positive incentives to ensure the support of the profession.

By contrast with general practice data, there are several technical challenges to overcome for hospital, medicines and laboratory data. The necessary enhancements to hospital data require changes in the way in which data are captured at the point of care, to enable more granular coding and more timely provision of data into national systems. This would be facilitated by the widespread adoption across hospital trusts of technology solutions to automate clinical coding. National data capabilities for hospital prescribed and dispensed medicines and for high-cost medicines are urgently needed to complement the data on community dispensed medicines and so provide a comprehensive picture of prescribed and dispensed medicines in England (section 3.1.5). This should be relatively straightforward since proof-of-concept demonstrations of these additional capabilities already exist.

The technical challenges for laboratory data are greater. They relate to the complexity of computer systems used to handle these data in laboratories across the country, wide variation in data formats and terminologies, and the very large numbers of tests requested (see section 3.1.6). But they are not insurmountable, provided initial efforts are focused on the relatively small number of tests (for example blood cell counts and biochemistry measures) that account for a very high proportion of test requests. Overcoming these technical challenges will rely on robust leadership, priority setting and addressing NHS England's capacity gaps.

Specialist data – unrealised potential

To realise their full potential, specialist audit, registry and screening datasets need to be accessible and linkable through a single national data access system. These datasets already exist as national collections, largely in structured, coded format. Despite this, the task is complex because of the very large number of these data collections, each with its own combination of data governance, management and curation challenges (section 3.1.12). Our recommendations in Chapter 7 involve addressing legal and regulatory issues relating to access, data handling and information governance capacity, and rigorous prioritisation of the work needed against several practical criteria.

At-scale access to and linkage of unstructured NHS imaging data brings a different set of technical and information governance challenges. England-wide scale will come from building on existing infrastructure investments and applying the experience, learning and expertise of pioneering radiology and pathology imaging data initiatives that have demonstrated success for populations of 5–10 million people (section 3.1.7).

Data from beyond the NHS

The Office for National Statistics (ONS) collects a wide range of data from non-NHS sources, including, for example, census, education and justice data. Many are health-relevant and, with the appropriate approvals, can be accessed for analysis and research via the ONS Secure Research Service or Integrated Data Service. Through the leadership of the Longitudinal Linkage Collaboration (LLC), several of these non-NHS data relevant to health and wellbeing will soon be linked to population-based longitudinal research cohorts held within the LLC's accredited trusted research environment (section 5.4). This work will also inform and support the linkage of these data to other large population-based cohorts such as UK Biobank and Our Future Health. Some linkages between healthcare data and data from non-NHS sources at whole-population scale in England have been achieved, but this has been tortuous, time-consuming and difficult. While the legal gateways provided in the Digital Economy Act 2017 (DEA) have enabled the sharing, linkage and analysis of data from public authorities for public good, the exclusion of NHS health and social care data from the DEA means that it has not enabled cross-sectoral linkages that include these data.[316] Revision of the DEA to include NHS health and social care data may be one potential solution. However, it would not be completely straightforward, for a couple of reasons. First, the DEA permits but does not mandate data sharing, and there would be a risk of potentially unhelpful overlap with the provisions of the National Health Service Act 2006 and the COPI Regulations, which provide a mechanism for the Secretary of State to mandate the sharing of NHS health and social care data. In addition, consultation, especially with health professional groups, would first be required, as would resolving how NHS data opt-outs would be handled.[317] Close partnership working between NHS England and the ONS to agree definitive, streamlined solutions for efficient and secure data access and sharing between their secure environments is also needed. These potential solutions are incorporated into our recommendations in Chapter 7.

316 See https://www.gov.uk/government/publications/digital-economy-act-2017-part-5-codes-of-practice/mid-point-report-on-use-of-the-dea-powers.

317 See: https://www.bma.org.uk/media/1619/bma-consultation-response-national-data-strategy-jul-2019.pdf.

Chapter 7

Recommendations and Conclusions

In this chapter

In this final chapter, we draw together a set of recommendations based on our review of public, patient and health professional views on health data uses (Chapter 2); our overview of the UK-wide health data landscape (Chapters 3-5); our summary of data and system needs and priorities of multiple diverse stakeholders (Chapter 6); and our consideration of barriers to the secure use of health data for public and patient benefit as well as potential solutions to these (Chapter 6). The recommendations focus on England but the principles behind them are relevant across the four nations of the UK. And some of our recommendations depend on UK-wide systems or strategies.

Boosting our use of data across the UK will bring huge opportunities: to improve health, wellbeing and economic productivity across society; to identify and fix inequalities; and to attract new and increased investment in the health, social and life sciences sectors in the UK. There are many barriers to be overcome and no single magic bullet solution. Rather, several complementary political, financial and technical solutions are needed. Multiple government, health, science, and data organisations must commit to working together to achieve cultural change; to reducing complexity across many dimensions; and to supporting ongoing, meaningful engagement with the public, patients, and health professionals. There is no room for institutional, organisational or other siloes, and those who believe they or their organisation can provide all the solutions are likely to hold back progress.

Ambition and vision are crucial, but must be accompanied by solutions that are practical, feasible, scalable and affordable. Innovative solutions must not be stifled, but the temptation to establish (potentially costly) new initiatives, instead of de-complexifying and de-duplicating what already exists, should be resisted. Strategic investments, for example to help plug capacity gaps, are important and necessary, but simply throwing more money at the barriers will not help on its own. Rather, aligning around common goals and putting in place incentive frameworks that focus the energies and resources of existing organisations will maximise success.

Government, health, science and data organisations should learn from the best examples of what already works well, given the remarkably successful health data-driven flagship initiatives in the UK that have informed healthcare and public health policy and improved health here and abroad. Some, such as UK Biobank, existed for many years before the COVID-19 pandemic, while others, such as the RECOVERY trial of COVID-19 treatments or new population-wide health data initiatives, emerged during – and as part of the response to – the pandemic. National public organisations need to recognise and apply advances made during the pandemic rather than simply returning to pre-pandemic business as usual. For example, during the pandemic, multiple national organisations aligned and worked together, enabling faster, broader access to data, delivering insights without data security or privacy breaches. Identifying the common goals to reinvigorate this collaborative, cooperative, efficient and proportionate way of working stands to benefit all of us.

7.1 System-wide recommendations

We make five recommendations for system-wide reform. These focus on England, in line with what the commissioners of this review requested. While the first three recommendations focus on England, their principles apply across all four nations. The last two recommendations are UK-wide.

1. Major national public bodies with responsibility for or interest in health data should agree a coordinated joint strategy to recognise England's health data for what they are: a critical national infrastructure, necessary to drive the generation of insights to maintain and improve health, wellbeing, safety, security and economic productivity (see also section 6.3.1).
2. Key government health, care and research bodies should establish a national health data service in England with accountable senior leadership, a ring-fenced budget and performance metrics, to accelerate research, analysis and innovation that benefits society.
3. The Department of Health and Social Care should oversee and commission ongoing, coordinated, engagement with patients, public, health professionals, policymakers and politicians, and should involve the public, patients and health professionals in how health data are used.
4. The health and social care departments in the four UK nations should set a UK-wide approach to streamline data access processes and foster proportionate, trustworthy data governance to enable more and better health data analysis, research and innovation for public and patient benefit.
5. National health data organisations and statistical authorities in the four UK nations should develop a UK-wide system for standards and accreditation of secure data environments (SDEs) holding data from the health and care system to accelerate the safe use of health data for research.

For each of these we provide more specific detail, outlining whether the need is political, financial and/or technical. We suggest which organisation(s) could take on primary responsibility and which others should be involved (**Who should be involved?**). We also outline what needs to be delivered (**What is needed?**) and suggest broad timelines for delivery (**By when?**). We refer to several national organisations, whose key functions and abbreviations are shown in Box 7.1.

Box 7.1 National organisations in England interested in or responsible for health-relevant data

Department of Health and Social Care (DHSC). DHSC is a ministerial department, supported by several agencies and partner organisations, including MHRA, UKHSA, NHSE, NICE, HRA, NDG (all covered below). DHSC is responsible for overall health and care policy in England and works with the devolved administrations of Northern Ireland, Scotland and Wales on UK-wide health and care priorities. It supports and advises ministers on health and social care policy, ensuring that the department and its arm's length bodies deliver on their commitments. It also plans for future domestic and global health needs; maintains and aligns legal, financial, administrative and policy frameworks; and acts to resolve complex, emerging health and care challenges.

NHS England (NHSE). As an arm's length body of the Department of Health and Social Care (DHSC), NHSE provides national leadership and oversight for the national health service in England. It supports and oversees the commissioning and delivery of health services by integrated care boards, allocating funding from the DHSC. It runs the national IT systems which support health and social care. It works with the wider NHS and partners to optimise the use of digital technology, research and innovation, and to deliver value for money and increased productivity and efficiency. It supports the collection, analysis and publication of – and access to – data generated by health and social care services to improve outcomes for patients.

National Institute for Health and Care Research (NIHR). NIHR is funded by the DHSC to fund, enable and deliver health and social care research that improves people's health and wellbeing and promotes economic growth. It works in partnership with the NHS, universities, local government, other research funders, patients and the public. It is also a major funder of applied health research in low- and middle-income countries (mainly via UK government international development funding).

UK Health Security Agency (UKHSA). UKHSA is an executive agency, sponsored by the DHSC. Established in April 2021, UKSHA merged the former roles of Public Health England (where these roles related to infectious diseases, chemical, radiation and environmental health threats) and two bodies established during the pandemic: NHS Test and Trace and the Joint Biosecurity Centre. UKHSA prevents, prepares for and responds to infectious diseases and environmental hazards, aiming to keep communities safe, save lives and protect livelihoods. It provides scientific and operational leadership, working with local, national and international partners to protect the public's health and build the nation's health security capability.

Health Research Authority (HRA). HRA is an arm's length body of the DHSC. Its core purpose is to protect and promote the interests of patients and the public in health and social care research. It ensures that research is ethically reviewed and approved; coordinates and standardises research regulatory practice; and provides independent recommendations on the processing of identifiable patient information where it is not always practical to obtain consent. HRA's functions apply mainly to research undertaken in England, but it works closely with the devolved administrations to provide a UK-wide system.

Medicines and Healthcare products Regulatory Agency (MHRA). MHRA is an executive agency sponsored by the DHSC. It regulates medicines, medical devices and blood components for transfusion in the UK. It is responsible for ensuring safety, quality and efficacy – and securing the safe supply – of medicines, medical devices and blood components; promoting international standardisation and harmonisation to assure the effectiveness and safety of biological medicines; educating the public and healthcare professionals about the risks and benefits of medicines, medical devices and blood components; enabling innovation and research and development to benefit public health; and collaborating with partners in the UK and internationally to enable access to safe medicines and medical devices and to protect public health.

National Institute for Health and Care Excellence (NICE). NICE is an executive non-departmental public body, sponsored by the DHSC. It provides evidence-based guidance on health services, social care and public health. This includes recommendations, in technology appraisals and highly specialised technologies guidance, on whether medicines and other treatments represent a clinically- and cost-effective use of NHS resources. NICE also produces clinical guidelines, public health guidance and quality standards.

National Data Guardian (NDG). The NDG is in independent public body, sponsored by the DHSC. It advises the health and adult social care system in England to help ensure that people's confidential information is kept safe and used properly. It aims to safeguard trust in the confidentiality of the health and social care system; support understanding and engagement about how and why data is used; and encourage safe, appropriate and ethical use of data in individual care, system planning, research and innovation that benefits the public. The NDG has the statutory power to issue official guidance about the processing of health and adult social care data in England but also provides informal advice.

NHS Business Services Authority (NHS BSA). The NHS BSA is an arm's length body of the DHSC, which delivers a range of national services to NHS organisations, NHS contractors, patients and the public. These include platforms and services to support the NHS workforce, primary care services including community pharmacies and dentists, and support for members of the public in accessing healthcare services and help with healthcare costs.

Health Quality Improvement Partnership (HQIP). HQIP is an independent organisation led by the Academy of Medical Royal Colleges and the Royal College of Nursing. It works on behalf of NHS England and other healthcare departments and organisations to commission, manage, support and promote national and local programmes of quality improvement. These include the national clinical audit programmes, local audit support resources and the National Joint Registry.

Department for Science, Innovation and Technology (DSIT). DSIT is a ministerial department supported by several agencies and public bodies, including UKRI (see below). It focuses on improving people's lives by maximising the potential of science and technology. It is responsible for positioning the UK at the forefront of global scientific and technological advancement; driving innovations that benefit people and the economy; delivering talent programmes, infrastructure and regulation to support the UK's economy, security and public services; and providing funding for research and development.

UK Research and Innovation (UKRI). UKRI is a non-departmental public body sponsored by DSIT. It invests in research and innovation to enrich lives, drive economic growth, and create jobs and high-quality public services across the UK. It supports researchers to develop new skills to further their careers; enables collaboration and engagement across research communities and the wider public; and invests in capabilities across the research system, including research infrastructure and an inclusive, ethical research culture. UKRI is made up of seven research councils, Innovate UK and Research England.

Health Data Research UK (HDR UK). HDR UK is the UK's national institute for health data science. It works to unite the UK's health data to enable discoveries that improve people's lives. Its vision is that every health and care interaction and research endeavour will be enhanced by access to large-scale data and advanced analytics. It is an independent charity organisation supported by nine funders, with work based at multiple locations across the UK. Its work spans across academia, healthcare, industry, and charities, as well as patients and the public.

Administrative Data Research UK (ADR UK). ADR UK is a major investment by the Economic and Social Research Council, part of UKRI. It is a UK-wide partnership, working to transform public sector data into research assets and policy-relevant insights. It does this by joining up the wealth of administrative data created by government and public bodies across the UK and making it available to approved researchers in a safe and secure way to support evidence-based policy decisions and more effective public services.

Office for Life Sciences (OLS). The Office for Life Sciences supports the delivery of the UK government's life sciences and innovation strategy by connecting decision-making across government. It champions research, innovation and the use of technology to transform health and care services, aiming to improve patient outcomes and support economic growth. OLS is part of the DHSC and DSIT.

UK Statistics Authority (UKSA). The UKSA is an independent statutory body which promotes and safeguards the production and publication of official statistics that serve the public good. It operates at arm's length from government as a non-ministerial department and reports directly to the UK Parliament, the Scottish Parliament, the Welsh Parliament and the Northern Ireland Assembly. The UKSA oversees the independent accreditation of processors, researchers and research projects that access or process data for research purposes under the auspices of the Digital Economy Act 2017.

Office for National Statistics (ONS). ONS is the UK's national statistical institute and its largest independent producer of official statistics. ONS is responsible for collecting, analysing and publishing statistics about the UK's economy, population and society at national, regional and local levels. It also conducts the census in England and Wales every 10 years.

Association of Medical Research Charities (AMRC). AMRC is a membership organisation dedicated to supporting medical research charities in saving and improving lives through research and innovation. It helps its member charities fund the best research by developing guides, providing training and auditing funding processes. It aims to drive positive change in the research and health landscape by influencing policy and research and by highlighting the sector's contribution to patient and public health.

Association of British Pharmaceutical Industry (ABPI). The ABPI is a UK organisation that represents companies of all sizes that invest in making and discovering medicines and vaccines to enhance and save lives. It aims to make the UK the best place in the world to research, develop and access medicines and vaccines to improve patient care. As part of its remit, it is committed to supporting UK policymakers and the NHS to enable efficient and legitimate health data access for research and care.

Association of British HealthTech Industries (ABHI). The ABHI is the UK's leading industry association for health technology, representing small, medium and large multinational companies that supply products from syringes and wound dressings to robots, diagnostics and digital technologies. These companies collectively play a key role in the delivery of healthcare and contribute significantly to the UK economy. The ABHI represents the health technology industry to stakeholders, including the government, NHS and regulators.

BioIndustry Association (BIA). The BIA represents the UK's innovative life sciences and biotech industry, supporting companies to start and grow successfully and sustainably. It has over 600 members, including start-ups, biotech and innovative life sciences companies, pharmaceutical and technological companies, universities, research centres, tech transfer offices, incubators and accelerators, and a wide range of life science service providers such as investors, lawyers and intellectual property consultants. It works across a wide range of areas including policy, finance, science, regulation, legal issues and talent.

7.1.1 Developing a coordinated, joint strategy to make England's health data a critical national infrastructure

The NHS is of course central in the generation of health data but, as we have seen, many other sources of health-relevant data exist, and the truly transformational insights emerge when they are linked together. A national strategy should foster easier and greater use of health data to benefit society. Many major, national, publicly funded organisations have a crucial role to play in developing a joint strategy because they generate, collect, manage, curate, fund or use data, with the aim of improving people's lives. In England, these include NHS England (NHSE), the Department of Health and Social Care (DHSC), the Department of Science Industry and Technology (DSIT), UK Research and Innovation (UKRI) and its constituent research councils, the National Institute for Health Research (NIHR), the Office for Life Sciences (OLS), the Association of Medical Research Charities (AMRC) and its constituent research charities, national research institutes/organisations (in particular Health Data Research UK [HDRUK] given its national remit and health data focus, but also Administrative Data Research UK [ADRUK]), the UK Health Security Agency (UKHSA), Office for National Statistics (ONS), Health Research Authority (HRA), Medicines and Healthcare products Regulatory Agency (MHRA) and the National Institute for Health and Care Excellence (NICE) (see Box 7.1).

Who should be involved?

DHSC, DSIT and ONS would be appropriate bodies to lead on drawing up an agreement outlining several commitments, involving all relevant national governmental and non-governmental organisations.

What is needed?

The need here is primarily **political**. We recommend that all relevant national organisations should sign up to the proposed agreement with the following commitments:

1. Acknowledge joint responsibility and accountability for reducing ecosystem complexity and fragmentation (with its organisational, transactional, computer system, legal and regulatory dimensions).
2. Commit to coordinated long-term planning and investment in publicly funded health data infrastructure, not driven by crisis management or unrealistic expectations on delivery timelines.
3. Support and contribute as appropriate to: the establishment and delivery of a national health data service to support health data-enabled research and analysis; nationally coordinated engagement with – and involvement of – patients, public, health and care professionals, policymakers and politicians in health data uses; a UK-wide approach to data access processes and governance; and a UK-wide system for secure data environment (SDE) standards and accreditation, (specified further in sections 7.1.2-7.1.5).
4. Establish mechanisms to ensure these commitments are met.

By when?

We recommend aiming to draw up and have this agreement signed within the first few months of 2025.

7.1.2 Establishing a national health data service in England with senior accountable leadership

A new national health data service should focus on uses of data beyond the delivery of individual patient care and the operational requirements of the NHS, such as waiting list and appointments management. Areas of work that the service should support include the mapping and understanding of population health needs, designing and evaluating screening and vaccination programmes, improving and implementing systems to monitor the safety of medicines and medical devices, and enabling clinical and population-based observational research studies and clinical trials. The service should recognise that analysts and researchers leading and contributing to these efforts come from NHS, academic, commercial and charity settings, often working in partnership across organisations. Such partnerships should be welcomed and facilitated. For maximum patient and public benefit, the national health data service must fulfil the needs of all these uses and users.

Who should be involved?

- The service should be overseen primarily by NHSE, NIHR, DHSC, DSIT and UKRI. These organisations should establish arrangements for joint accountability for the service. Delivery through existing organisational structures would avoid the complexity of establishing a new body. The service would benefit from being delivered via a strategic partnership with NHS, academic and industry users.
- Governance arrangements should include an advisory board with representation from NHSE, UKRI, NIHR, OLS, ONS, MHRA, DHSC, NICE, UKHSA, AMRC, NHSBSA, HQIP, HDR UK, ABPI, ABHI, and BIA (see Box 7.1 for explanation of these abbreviations).
- Senior accountable leadership of the service will be crucial. It should be led by a senior executive director, with credibility among clinical, research, policy, and data science/technical expert communities, and the ability to drive cultural change. We suggest that they should report directly and jointly to the CEO of NHSE, the Government Chief Scientific Adviser for Health/CEO of NIHR, and the Executive Director of UKRI. They should have both responsibility for delivery of – and ring-fenced budget for – the national health data service.

What is needed?

Political leadership will be needed both to establish the service and to support some of the legislative changes that may be required to enable data access. The service could potentially be set up within existing organisational structures, provided there is clarity on joint accountability from the key leading organisations (notably NHSE, DHSC/NIHR and DSIT/UKRI), senior leadership, a ring-fenced budget, and agreed performance measures. There are **financial** implications as the delivery of the service itself will require ring-fenced government investment. However, much of this should come from redirecting existing investment via a new and more efficient delivery model. Delivery of the service will also need to overcome a range of **technical** challenges, such as the implementation of data services and dataset provision.

A new national health data service should undertake the following tasks:

1. **Drive cultural change within and across NHS England** to become an organisation that **understands, supports, champions and gains from a wide range of research and analysis** using the data it generates.

2. Deliver through **strategic partnerships with the health data user community** (including **NHS, academic and industry-based users**) to provide expert input and enable innovative, streamlined, user-informed system design.

3. Establish and oversee a **single national health data access system for England** with **streamlined and standardised data governance and access**, modelled on the Integrated Research Application System, the system for permissions and approvals of health and social care research in the UK. The single national health data access system should include **performance monitoring, targets and incentives** that maximise beneficial uses of data, learning from successful data access processes that have scaled well (for example UK Biobank). It should support access **within national and regional SDEs** as well as **data transfer to other secure locations** where appropriate.

4. Develop a **practical plan with the devolved nations and the ONS Integrated Data Service for cross-nation and cross-sectoral data sharing/access and linkage** and input into a **joint UK-wide approach for data access processes and governance, and on standards and accreditation for SDEs** (see 7.1.3 and 7.1.4). Partnership with the ONS and UKSA is key to ensuring access to and linkage of data from administrative health-relevant sources beyond the healthcare system, while partnership across the four nations of the UK is necessary for analysis and research efforts with UK-wide reach.

5. **Address capacity and capability gaps**, especially within NHS England, including through:

   - rebuilding and enhancing capacity in **specialist health data information governance** and in **data management and curation**;

   - working with delivery partners to facilitate this, for example through **secondments into NHSE** to address headcount and specialist expertise constraints – for example from HDR UK, universities, DHSC, industry, UKHSA and others.

6. Implement **practical, acceptable and transparent data infrastructure investment strategy and data access cost recovery and pricing models** (moving beyond the principles and policy so far produced to practical implementation). These should incorporate:

   - **strategic precompetitive industry and philanthropy investment models**, for example drawing on the success of UK Biobank in attracting this type of inward investment; and

   - **transparent, well justified pricing models for public sector, non-profit and for-profit uses.**

7. Lay out and implement a clear **roadmap for data services and dataset provision,** including:

   - coherent, logical integration of new national data infrastructures and services (for example the NHSE and regional SDE network, OpenSAFELY, NHS DigiTrials) with pre-existing ones (for example NHSE Data Access Request Service, Clinical Practice Research Datalink (CPRD));

   - priorities for incorporating and enabling access to national generic and domain-specific data assets, with ambitious but realistic timelines;

- data services, which should include:
  - improved metadata;
  - data quality improvement strategies (including encouraging wider use to drive quality);
  - standardisation of data collection, formats and terminologies;
  - reproducible data management, curation and code development processes;
  - enhanced person- and place (Unique Property Reference Number)-level linkage capabilities (including descriptions of linkage methods and record-level match quality indicators);
  - improved user experience, for example through data curation and analysis support, mandatory sharing of protocols, code and algorithms, required adherence to practices to increase efficient use of shared compute, and ability to support advanced analytic approaches, including AI;
  - integrate, enhance and expand NHS DigiTrials and CPRD 'find, recruit and follow' services for trials and longitudinal cohorts, including whole-population general practice data.

8 Clear plans for **development of complementary national and regional data infrastructures**, which should:

- ensure that NHS local and regional organisations can rapidly access and benefit from national data assets relevant to their own regional planning, research and innovation;
- support consistent and logical development of regional data infrastructures (for example regional SDEs and NIHR regional infrastructure investments) that add complementary detail and granularity of data to existing and future planned national infrastructure and data;
- contain resource needs by avoiding duplication across national and regional infrastructures.

By when?

Suggested timelines that would demonstrate appropriate intent and ambitious goals for the national health data service are as shown below:

Task	Delivered by
Leadership, advisory board and strategic delivery partnership in place	End Q2 2025
Roadmap for data services and dataset provision, and clear plans for complementary national and regional data infrastructure	End Q3 2025
Single national health data access system launched and functional, with initial targets set, monitored and reported (for example completed applications assessed within two weeks of submission, definitive decision within one month, aiming for >90% approval through clear guidance for applicants, data access within two months)	End Q4 2025
Solution to address capacity gaps developed and implemented	End Q3 2025
Data infrastructure investment strategy developed	End Q3 2025
Transparent cost models introduced	End Q3 2025
Practical plan for data access/sharing/transfer/linkage mechanism between NHSE and devolved nations (for seamless four nations analyses), and between NHSE and ONS (for health to non-healthcare data linkages)	End Q4 2025

7.1.3 Ongoing, coordinated engagement with patients, public, health professionals and policymakers

Who should be involved?

- Given its ongoing work in this area, the DHSC is likely to be the most appropriate body to oversee and commission this engagement.
- Two organisations that could play a very helpful role in jointly convening, coordinating and advising nationally are:
  - **Understanding Patient Data,**[318] which has a longstanding track record of providing clear, transparent and consistent information – focusing on data from the health and care system – of relevance to patients, public, health professionals and policymakers across the four nations of the UK; and
  - the **Public Engagement in Data Research Initiative (PEDRI),**[319] a cross-sectoral partnership that has: ongoing engagement activities and aims to promote good practice relevant to multiple sources of data from outside as well as within the health and care system; a UK-wide remit; joint leadership from ADRUK, Cancer Research UK, DARE UK, HDR UK, NHS England, ONS, the Office for Statistics Regulation, Research Data Scotland, Smart Data Research UK and the UK Longitudinal Linkage Collaboration; and wider partnerships.

Other relevant organisations should also be involved. These include many of the national organisations listed in section 7.1.1, several of which have made a shared commitment to embed public involvement in health and social care research,[320] the National Data Guardian, and other patient, public or professional facing organisations (such as Use My Data, National Voices, medConfidential, the British Medical Association and the Academy of Medical Royal Colleges).

What is needed?

This initiative will require **political** and some **financial** support. It should build on the existing large-scale public engagement efforts that NHSE and DHSC are leading,[321] broadening these to include engagement with and meaningful involvement of health professionals and policymakers and to encompass health-relevant data from beyond the health and care system.

It should include the following:

1. Consistent narrative from relevant national organisations delivered in different ways to resonate with multiple audiences, focusing on the health, wellbeing and economic[322] benefits for all patients, public and health professionals from a wide range of data uses.
2. A multi-pronged, multi-organisational strategy for ongoing engagement with multiple segments of society about how data can help solve not only COVID-19-related but also non-COVID-related healthcare and public health challenges.

318 See https://understandingpatientdata.org.uk/.

319 See https://www.hdruk.ac.uk/news/introducing-the-public-engagement-in-data-research-initiative/.

320 See https://www.hra.nhs.uk/planning-and-improving-research/best-practice/public-involvement/putting-people-first-embedding-public-involvement-health-and-social-care-research/.

321 See https://transform.england.nhs.uk/key-tools-and-info/data-saves-lives/national-public-engagement-on-the-use-of-health-data/.

322 Economic gains come from increased health and wellbeing via increased productivity across all groups (by age, ethnicity, geography, deprivation etc).

3. A major focus on understanding the perspectives of health professionals, especially GPs, given the high priority placed on access to data from general practice systems.

- Acceleration of efforts for patients to be able to access to their own health data, noting:
- the increasing uptake of the NHS App, although recognising that this route is not suitable for everyone;
- that trust in data use and accuracy of health data will be increased if people can easily access their own health data;
- that practical steps are needed to broaden the opportunity without placing unnecessary burden on busy healthcare professionals (especially GPs).

5. Accelerate and inform a single, consistent, centralised, readily accessible system in England for NHS data access opt-outs that does not impose any burden on busy GPs.

By when?

(1), (2) and (3) are ongoing activities

For (4) and (5), given pre-existing and ongoing activities, it should be possible to generate policy advice and plans for rapid implementation by Q3 2025.

7.1.4 Setting a UK-wide four nations approach for data access processes and proportionate data governance

Who should be involved?

Setting this approach will require **political** leadership and engagement from across the four nations. This should be led by the DHSC and health and social care departments of the devolved administrations. Coordination by a UK-wide organisation will be needed. This could be provided by the Pan UK Data Governance Steering Group of the UK Health Data Alliance,[323] which has a UK-wide remit and broad membership and is already leading work in this area. The Office for Strategic Coordination of Health Research (OSCHR) could potentially play an additional coordinating role,[324] given its UK-wide remit and representation. Input will also be needed from NHSE and NHS bodies in the devolved administrations, the Health Research Authority, National Data Guardian, Understanding Patient Data, data users in academia, industry and NHS, ONS, UKSA, the Information Commissioner's Office (ICO) and others.

What is needed?

The approach should aim to build on and align policy developments across the four nations.[325]

The approach should also confront legal and regulatory complexity by:

1. Providing clear and transparent guidance for data providers and users on current approaches for accessing health data, including the legal and regulatory requirements.

323 See https://ukhealthdata.org/projects/data-access-and-governance/ and https://ukhealthdata.org/alliance-outputs/.

324 See https://www.nihr.ac.uk/about-us/what-we-do/working-with-partners/other-funders/oschr.

325 See https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data (England); https://www.gov.scot/publications/data-strategy-health-social-care-2024-update-progress-priorities/ (Scotland); https://www.gov.wales/digital-and-data-strategy-health-and-social-care-wales-html#127707 (Wales); https://dhcni.hscni.net/wp-content/uploads/2023/06/doh-hscni-data-strategy.pdf?csrt=6744512751555423319 (Northern Ireland).

2. Proposing governance frameworks that focus on realising the benefits of using health data (rather than solely on minimising risk), that will enable streamlined cross-sectoral linkages and that reduce unnecessary differences in approaches for accessing data across the UK.

3. Recommending where new or revised legislation will help as well as where clear, consistent interpretation of current legislation and common law requirements would be as effective. Priority areas include the following:

- lay out rationale and mechanism to extend legal gateways successfully applied during the COVID-19 pandemic to a much broader set of health-related uses of data with pressing needs;
- lay out the pros and cons of including health and social care data in the Digital Economy Act;
- work with UK-wide population- and disease-based cohorts[326] to clarify and recommend legal gateways for data access and linkage.

By when?

- Guidance on current approaches by the end of Q2 2025.
- Proposed improvements and recommendations by the end of Q2 2025.

7.1.5 Developing a UK-wide system for standards and accreditation of SDEs holding data from the health and care system

A UK-wide system for standards and accreditation of SDEs will accelerate the safe use of health data across the UK for patient and public benefit. It will require **political support** and ongoing **financial support** for key organisations involved in coordination and delivery.

Who should be involved?

The UK Statistics Authority,[327] working together with the health and social care departments in the four UK nations, would be well positioned to lead on SDE accreditation, given its UK-wide remit, areas of focus, and existing role in accrediting secure processing environments under the Digital Economy Act (DEA).[328] HDR UK, ADR UK, the UK Health Data Research Alliance,[329] and UKRI's Data and Analytics Research Environments UK (DARE UK) programme[330] would be well positioned to lead on SDE standards given their UK wide remit and existing work in this area. Input and buy in would also be required from SAIL Wales, the Northern Ireland Trusted Research Environment (NITRE), Research Data Scotland, NHS England (in particular its Data for Research and Development programme), and the ONS Secure Research Service and Integrated Data Service teams.

326 In particular via cohort collaborative organisations, such as Population Research UK (https://www.ukri.org/what-we-do/browse-our-areas-of-investment-and-support/population-research-uk/), Longitudinal Linkage Collaboration (https://ukllc.ac.uk/), BHF Data Science Centre disease-based cohorts platform (https://bhfdatasciencecentre.org/areas/cohorts/).

327 The UK Statistics Authority is an independent body at arm's length from government. It has a statutory objective of promoting and safeguarding the production and publication of official statistics that serve the public good. See https://uksa.statisticsauthority.gov.uk/.

328 See https://uksa.statisticsauthority.gov.uk/digitaleconomyact-research-statistics/better-access-to-data-for-research-information-for-processors/.

329 See https://ukhealthdata.org/.

330 DARE UK (Data and Analytics Research Environments UK) is a programme funded by UKRI to design and deliver coordinated and trustworthy national data research infrastructure to support cross-domain research for public good. See https://dareuk.org.uk/.

What is needed?

1. A **UK-wide SDE accreditation framework** for SDEs holding data from health and care system, which should:
   - build on established and widely supported schemes, in particular UKSA accreditation under the DEA;
   - extend recent work done by the DHSC and UKSA on an accreditation framework for NHS SDEs in England to develop a UK-wide SDE accreditation framework that will enable data access arrangements across all four nations of the UK (see section 5.5.2);
   - provide strong rationale for additional specific criteria for SDEs holding health and care data compared with other types of sensitive, personal data (for example the ability to apply opt-outs where relevant).
2. Recognised **UK-wide SDE standards**, building on the work of the Standard Architecture for Trusted Research Environments (SATRE) project (see section 5.5.3),[331] which provides guidance for SDEs across the areas of information governance procedures, computing technology, data management and various supporting capabilities, aiming to standardise the capabilities of SDEs. These should include:
   - agile adaptation to user feedback and user needs;
   - promotion and incentivisation of positive user behaviours (for example efficient coding, sharing of protocols, code and algorithms) that benefit all users.
3. Guidance and policy on avoiding an unhelpful excess of SDEs (which would add rather than reduce complexity).

By when?

Pre-existing relevant activities mean that (1), (2) and (3) should be possible by the end of Q3 2025.

7.2 Data-specific recommendations

In section 6.2 we summarised the key datasets and data types that are seen as high priority by a wide range of diverse stakeholders. And in section 7.1, we recommend that major, national health, care, and research bodies in England establish a national health data service to deliver a range of services that include laying out and implementing a clear roadmap for dataset provision. Here we provide recommendations to guide that roadmap, specific to different high priority data types and sources. Our recommendations fall into three main areas:

1. General practice data: there is a need to ensure secure access to comprehensive, coded general practice data at national, whole-country scale, in near real time when necessary and linkable to other data sources for the full range of beneficial uses.
2. Other major, national and regional health and care data assets: there is a need to prioritise and fix issues affecting access to data from hospitals, medicines data, laboratory data (including genomics), national audits and registries, screening data and unstructured clinical data (including imaging and free text).
3. Data from other sectors: there is a need to develop capability that will make access to national health-relevant data from other sectors and their linkage to national, health and care data sources 'business-as-usual' rather than 'by exception' activities.

Progress for each of these data categories will require a combination of **political, financial** and **technical** solutions.

331 See https://satre-specification.readthedocs.io/en/v1.0.0/.

7.2.1 Establish a national system for general practice data

Who should be involved?

This is the highest data priority for the proposed national health data service (section 7.1.2). It will need to involve the GP profession (represented by the RCGP and BMA), NHSE, DHSC and the primary care computer system suppliers.

There is an urgent need for progress. This cannot and should not wait for the establishment of the national health data service. Hence we recommend an interim solution to ensure progress is made based on existing capabilities while a more definitive solution is put in place. This interim solution should be led by DHSC, NIHR and NHSE. It will need engagement from the RCGP and BMA, ONS, England's National Data Guardian, the Health Research Authority, major research cohort and clinical trial leads (for example UK Biobank, Our Future Health, the Longitudinal Linkage Collaboration) and organisations representing the views of patients and the public (for example Understanding Patient Data).

What is needed?

There is an urgent need for researchers and policymakers to have more streamlined, broader access to whole-population, comprehensive, coded data from general practice computer systems, linked to other sources of health data. This was the most frequently raised issue during consultation for this review. Current mechanisms of access to general practice data do not enable the full range of beneficial uses. A single national centralised system, supporting both access to general practice data within the NHS England SDE as well as transfer of subsets of these data to other secure locations, with appropriate safeguards in place, would support the wide range of beneficial uses shown in Table 7.1.

Table 7.1 Benefits required of a national system for general practice data

1. Policymakers and NHS delivery teams can properly plan and equitably deliver healthcare for everyone, using general practice data from the whole population. For example:
   - assessing where the need for increased general practice resources and services is greatest;
   - accurately identifying people to invite for vaccination and screening programmes;
   - providing a data analysis service to fulfil GPs' needs for data about their own practice populations.
2. Researchers/analysts can undertake inclusive, whole-population research using multiple, linked sources of health data, including general practice data, securely analysed within a secure data environment, to generate discoveries and insights that inform better health and care for all health conditions.
3. Researchers/analysts can use linked general practice data for better characterisation and follow-up in research cohorts and clinical trials of engaged participants who consented to such linkage, with these data provided via secure transfer mechanisms already used for other national data such as hospital episode statistics.
4. Researchers/analysts can access whole-population general practice data linked to non-healthcare data at whole-population scale (for example within the ONS secure setting) to allow key policy-relevant questions to be addressed, such as understanding the causes of long-term sickness in the 2.8 million working age people in England currently not working due to health problems, and how they might be supported.
5. Researchers/analysts can use whole-population general practice data, linked to other national health data sources, to inform centrally coordinated invitations (for example via NHS DigiTrials) to people eligible to participate in research relevant to a wide range of health conditions, giving as many people as possible the opportunity to take part.

Rapid interim solution

The need here is primarily **political**. We recommend two key actions to enable rapid progress towards the uses of general practice data for benefits shown in Table 7.1:

1. Rapid implementation of Secretary of State directions that enable the NHS England General Practice Data for Pandemic Planning and Research data to be used for research and analysis for non-COVID-19 health conditions.
2. Accelerate the planned extension of the Secretary of State direction for research and analysis using the OpenSAFELY platform within the TPP and EMIS primary care computer system suppliers' data centres to cover non-COVID-19-related health conditions.

The NHS England General Practice Extraction Service Data for Pandemic Planning and Research dataset (GDPPR) – a population-wide general practice data extract covering 98% of English general practices – was established in the early months of the pandemic.[332] This dataset continues to flow regularly from general practice computer system suppliers to NHS England's secure systems under direction from the Secretary of State for Health for COVID-19 research and analysis purposes. It includes a large subset of structured coded data (including many of the most extensively used codes) on almost every person in England. GDPPR data are already accessible, linked to other health data, within the NHSE SDE. With the appropriate approvals,[333] GDPPR data have also been transferred from NHSE to specific secure external locations. Examples include the ONS, which has received GDPPR data as part of the ONS Public Health Data Asset, and the University of Oxford coordinating centre for the RECOVERY trial of COVID-19 treatments, which has received GDPPR data on trial participants. As a result, these data have been and continue to be used to generate COVID-19 related insights for patient and public benefit.

Since the necessary secure data flows and approvals processes are already established, Secretary of State directions that extend the uses of these data to research and analysis for non-COVID health conditions would very quickly extend the patient and public benefits. We recommend the rapid implementation of directions that will allow the GDPPR data to be used for all the purposes shown in Table 7.1.

Plans were announced in November 2023 to extend the current Secretary of State direction for COVID-19-related research and analysis using the OpenSAFELY platform within the TPP and EMIS primary care computer system suppliers' data centres to cover non-COVID-19-related health conditions.[334] However, a year later, this extended direction has still not been implemented. While the current configuration of the OpenSAFELY platform does not support the full range of uses (Table 7.1), it does enable analyses using all coded information in the general practice records (rather than a subset, as for GDPPR). The planned, extended direction will substantially expand the range of inclusive, population-wide research and analysis that OpenSAFELY can support, and its implementation should be accelerated.

332 See https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research/guide-for-analysts-and-users-of-the-data.

333 Via NHS England's Data Access Request Service (DARS), with oversight from the independent Advisory Group for Data (https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/advisory-group-for-data/who-are-agd-and-what-do-they-do), with additional approval by BMA/RCGP representatives.

334 See https://www.england.nhs.uk/2023/11/nhs-expands-use-of-secure-covid-19-research-platform-to-help-find-new-treatments-for-major-killer-conditions/.

Definitive solution

The proposed interim solution to rapidly enable the broader use of general practice data for research and analysis would be a major step forward. However, the definitive solution needs to enable secure access to approved users from the NHS, universities and industry to fully comprehensive, national, structured, coded general practice data (rather than the subset available via the GDPPR) for each of the beneficial uses shown in Table 7.1. There are several potential options for this secure access, with their pros and cons explored in Appendix 11. There may be other viable options but, in brief, those considered here include:

1. Comprehensive, population-wide, structured, coded data extracted from general practice computer systems into NHSE systems.
2. Expand the Clinical Practice Research Datalink (CPRD).
3. Expand the RCGP Research and Surveillance Centre (RSC).
4. Implement OpenSAFELY capabilities within the NHSE SDE.
5. Explore data within general practice computer systems via OpenSAFELY and extract subsets of data to NHSE (or other secure settings) as needed.

Appendix 11 shows that option (1) is the only option able to deliver against each of the uses shown in Table 7.1. This could be delivered in partnership with CPRD, RCGP RSC and OpenSAFELY, so incorporating the advantages of options (2), (3) and (4). This could potentially be done through incorporating CPRD and RCGP RSC capabilities, services and expertise within the proposed national health data service and through implementing the OpenSAFELY platform within NHSE systems. Together these would enhance the reproducible data curation and analysis pipelines and the privacy and security features currently available within NHSE systems. A detailed consideration and delivery of one (or a combination of more than one) of these options is needed.

We recommend that the national health data service establish a national general practice data task force to oversee the appraisal of these options and delivery of the preferred option(s).

The need here is primarily **political** although some **technical** and **financial** support will be needed.

By when?

Interim solution – given the necessary technical solutions are already in place, it should be possible to implement this within the first few months of 2025.

Definitive solution – given the pressing need for a national general practice data solution we suggest:

- establishing a national general practice data task force within the first few months of 2025;
- completing the proposed options appraisal and agreement on preferred option(s) by the end of Q3 2025;
- full implementation during the early months of 2026.

7.2.2 Improve and accelerate access to other major national and regional NHS data assets

Who should be involved?

This should be overseen by the newly established national health data service (section 7.1.2).

What is needed?

A combination of **political, technical** and **financial** solutions will be required.

The national health data service needs to **understand the data priorities of a wide spectrum of data users and put in place a roadmap and actions to address these**. High priority data types and sources emerging from consultation for this review were discussed in section 6.2.

Beyond general practice data (dealt with in section 7.2.1), priority data fall into four broad categories:

1. **Structured, coded datasets that are already collected and collated nationally but that cannot be readily accessed or linked to other datasets via a centralised mechanism**. These include screening data, many national disease audit and registry datasets and adult social care data (see sections 3.1.8, 3.1.12 and 3.2.2). **We recommend that the national health data service should establish a national data task force to oversee the prioritisation, consolidation and accessibility of these data via the proposed single national health data access system** (section 7.1.2). Retaining and, where needed, expanding domain-specific knowledge to quality assure and curate these data will be important. Secretary of State directions will be required to enable the acquisition within NHS England of some of these data. Accessibility will also depend on increasing NHSE's specialist information governance and data management and curation capacity (section 7.1.2).

- Enabling access should be easiest for those **national datasets which are already acquired and controlled by NHS England**. All such datasets should be made available through NHSE's Data Access Request Service (DARS) for access within the NHSE SDE as well as for extraction and external secure transfer where necessary. This is not the case at present. For example, screening datasets, cancer registry, rare diseases and national diabetes audit data either require a separate application process or cannot yet be accessed within the national SDE. Person-level national adult social care data are now collected quarterly and should also be made available via DARS.
- Some national audit datasets (for example stroke and cardiovascular audits) are **commissioned by NHS England (either directly or through the Health Quality Improvement Partnership) and provided to NHS England but for COVID-specific purposes only**. For these, extended Secretary of State directions could be implemented to enable the much wider benefits that would arise from uses across all health conditions.
- Others, such as the National Respiratory Audit Programme, National Vascular Registry and Paediatric Intensive Care Audit Network are **commissioned by the Health Quality Improvement Partnership on behalf of NHS England but not currently provided to NHS England** to be made available via DARS. Secretary of State directions like those proposed for the stroke and cardiovascular audits are needed to allow this improved access.
- In addition, there are **national audits and registries not commissioned or controlled by NHS England** (for example the national renal registry and the out of hospital cardiac arrest audit) which could be hugely valuable if made available and linkable to other national data via NHS England's DARS.

The proposed national health data service will need to prioritise tasks to improve access to these datasets. This will involve considering several criteria, including:

- the **importance** of the health condition(s) addressed, for example with respect to incidence, prevalence, mortality, morbidity, or direct and indirect costs to health, care and society;
- **lack of adequate data from other accessible sources** about the health condition(s);
- **demand** from a wide range and/or large number of potential users (bearing in mind that volume of data requests may be a poor proxy for demand, as many potential users may not know about the data or how to request it, while others may not request access because of a belief that the access process will take too long or be unsuccessful);
- whether or not the data are **already held within NHS England** central systems or need to be obtained from an external organisation;
- the **capacity and willingness of the data controller and/or processor** to provide the audit/registry data for access and linkage by regular supply of updated data to NHS England;
- **ease of processing and curating the data centrally** prior to providing as a dataset for access and linkage (noting that some datasets are particularly well managed and curated and so less challenging to handle).

1. **National hospital episodes data, which need to be enhanced** to include more granular diagnostic and procedural codes and access in near real time. **We recommend that the national health data service establishes a national hospital data task force to oversee the adoption of approaches to achieving these enhancements as laid out in Appendix 10**. This will require changes both in the way data are acquired in clinical settings and in how they are provided centrally. These will take longer to implement than requirements for the already existing datasets covered in point (1).
2. **High priority data which are not yet collated at national scale**. These include data on **medicines** prescribed and dispensed in hospital, data on high-cost medicines and data on **laboratory** assay test requests and results. **We recommend that the national health data service establishes national medicines data and national laboratory data task forces to oversee the adoption of the approaches laid out in Appendix 10, building on existing expertise and demonstrations that these national data assets are achievable.**
3. **Unstructured data**, including radiology and pathology imaging data and data from the free text of electronic medical records. **We recommend that access at scale to these types of data in England should take advantage of existing expertise and infrastructure developments in tools and platforms for imaging and free text data, as well as investment in and increasing capabilities of regional SDEs.**

- For **imaging data**, as discussed in section 3.1.7 and Appendix 10, significant technical challenges must be overcome to establish national systems for sharing, viewing, reporting, accessing and analysing both radiology and pathology images. Such systems are already urgently needed for providing joined-up care, managing workforce challenges in radiology and pathology, and supporting large-scale analyses based on linking data from images to other sources of health data at national scale. National solutions for both radiology and pathology imaging data should build on existing specialist infrastructure and capabilities,[335] and integrate with NHS England's Data for Research and Development secure data environment (SDE) programme. **We recommend that the national health data service establishes national radiology and pathology imaging data task forces to plan and oversee the development of services for national-scale radiology and pathology imaging data access for research and analysis by no more than two to three of the 11 regional NHS SDEs.**

- For **free text data**, increasingly sophisticated solutions are being developed for enabling secure access to anonymised free text electronic medical records[336] and for extracting structured information from unstructured medical free text.[337] Such tools are being used across several hospital trusts. However, these are not yet widespread. Deployment of these tools across both primary and secondary care electronic patient record systems will increasingly enable access to the large proportion of health data held in unstructured free text format and its use to increase the efficiency, effectiveness and safety of individual patient care, service planning, and research. Structured data generated through use of these tools (for example via automated clinical coding) could be incorporated in national datasets. The free text itself is likely to remain within the original electronic patient record systems, although some may be pooled across hospitals within regional SDEs (for example collections of free text radiology or pathology reports). Now that general practices and almost all hospitals have electronic patient records, **we recommend that DHSC should invest in the efficiency gains of deploying tools to securely access and extract structured information from unstructured medical free text across all primary and secondary care EPR systems.**

335 For pathology imaging in England, these include the capabilities of the National Pathology Imaging Cooperative (https://npic.ac.uk/) and PathLAKE (https://www.pathlake.org/) and the digital pathology expertise of the Royal College of Pathologists (https://www.rcpath.org/profession/committees/digital-pathology-committee.html). For radiology imaging, building on the expertise of the London AI Centre for Value Based Healthcare, drawing on the learnings of large-scale imaging data pooling initiatives (see section 3.1.7), and aligning with the clinical and professional expertise of the Royal College of Radiologists (https://www.rcr.ac.uk/) will be important.

336 E.g. Clinical Record Interactive Search, see https://oxfordhealthbrc.nihr.ac.uk/about-us/core-facilities/cris/.

337 E.g. Cogstack, see https://cogstack.org/.

By when?

Suggested timelines for the national health data service to complete the recommended actions are as follows:

Task	Delivered by
Roadmap for dataset provision	End Q3 2025 (as per section 7.1.2)
Establish national task forces for screening, audits and registries, adult social care, hospital data, medicines, laboratory, radiology imaging and pathology imaging data	Mid 2025
All datasets that are already collected and collated nationally available via the single national data access system	End 2025
Enhancements to hospital episodes data in place	End 2026
National data on hospital and high-cost medicines available via the single national data access system	End 2025
National data on the 80-100 most common laboratory assays available via the single national data access system	Early months of 2026
Roadmap for national solution for long tail of remaining laboratory assays	Early months of 2026
Roadmap for access to radiology and pathology imaging data at national scale	Early months of 2026
Deploy tools at national scale for secure access to and automated extraction of structured information from unstructured medical free text across primary and secondary care EPR systems	Early months of 2027

7.2.3 Transform access to data from social care and other sectors linked to healthcare data at national scale

Who should be involved?

This will require committed and coordinated partnership working across several organisations, including the national health data service, DHSC, NHSE, ONS, UKSA and the Pan UK Data Governance Steering Group.

What is needed?

The need here is primarily **political**, with some **technical** and **financial** requirements.

Even during the pandemic, there were considerable difficulties in linking multiple sources of national health data collected and held by NHS England to health-relevant data collected from a wide range of sources and held by the ONS (section 3.2). Such linkages are essential to enable analyses of existing data to inform on the wider determinants of health, to assess and address inequalities, and to understand the links between healthcare and public health policies, health and wellbeing, and economic productivity. We know that such linkages are possible (for example the development of the ECHILD resource linking children's education data to health data from hospital episodes, or the ONS Public Health Data Asset, linking census data, hospital episodes data and general practice data for COVID-related research and analysis), but resource intensive and time-consuming. The aspiration is for access to national health-relevant data from other sectors and their linkage to national health and care data sources to be routine rather than exceptional activities.

In section 7.1.2, we recommended that the national health data service should lead on the development of a practical plan for data access and sharing between NHSE and the ONS that would enable such linkages to occur. Data sharing between the two organisations and between their respective SDEs would be facilitated by:

- regular communication between nominated senior representatives of the legal and information governance teams within the two organisations to foster trust and familiarity;
- a review of the pros and cons of including health and care data in the Digital Economy Act (see section 7.1.4);
- the implementation of a UK-wide system for accreditation and standards of SDEs holding data from the health and care system that NHSE and ONS SDEs could both sign up to (see recommendations in section 7.1.5).

By when?

Establish a regular tempo of communication at senior level between NHSE and ONS within the early months of 2025.

Provide recommendations on Digital Economy Act by end Q2 2025 (see 7.1.4)

Provide recommendations on SDE accreditation and standards by end Q3 2025 (see 7.1.5)

7.3 Concluding comments

In summary, progress will depend on cooperation, collaboration and cultural change across the relevant national public organisations, together with bold and visionary leadership. These organisations must commit to a joint strategy to make England's health data a critical national infrastructure to drive health, wellbeing and economic productivity. They must back a senior executive leader to oversee a new national health data service for England to enhance patient and public benefit through enabling the rapid, efficient, secure, national scale use of different sources of health data. They must also support ongoing, coordinated engagement with patients, public, health professionals, policymakers and politicians.

The recommendations focus on England. But there is also a pressing need for the joint initiatives, across the four nations of the UK. These should include a UK-wide approach to streamlined data access processes and proportionate data governance, with improved, more consistent use of existing legal gateways for data access and updated legislation where necessary. In addition, agreed UK-wide systems are needed for setting standards and accrediting secure data environments holding data from the health and care system.

With respect to data types and sources, the highest priority is to put in place a national system that enables secure access to comprehensive, coded general practice data for the whole population, linkable to other data sources and capable of supporting the full range of beneficial use cases. There is also a need to prioritise and fix issues affecting access to data at national scale from hospitals, medicines data, laboratory data (including genomics), national audits and registries, screening data, social care data and unstructured clinical data (including imaging and free text). Finally, close partnership working between NHS England and the ONS, supported by any necessary legislative change, is needed to ensure that access to and linkage of health-relevant data from other sectors becomes a routine rather than a cumbersome, bespoke process.

Implementing these recommendations will be tough but not impossible. For each barrier to be overcome, there are compelling examples to demonstrate what is possible. Success will be about ensuring that these can be replicated and scaled efficiently, so that they become the norm rather than the exception.

Appendices

Appendix 1
Review team

About the lead author

Professor Cathie Sudlow is, first and foremost, a medical doctor. She has almost three decades of clinical experience as a general physician, neurologist and stroke specialist in the National Health Service (NHS). She has worked in many different healthcare settings across the UK, both in large academic teaching hospitals and district general hospitals, including in general and specialist wards, intensive and coronary care units, emergency departments and outpatient clinics (including some based in community general practice settings).

Her research interests have always been firmly embedded in the world of 'big data' and underpinned by a commitment to improve the health and well-being of patients, their families and carers, and the wider public. As a scientist, she has broad expertise and experience in population health, clinical and genetic epidemiology, clinical trials and statistics. Her focus over the last 15 years has been on leading large-scale, interdisciplinary, collaborative, open-science initiatives that bring a better understanding of the causes and consequences of health and disease across the life course, leading to new and improved approaches to prevention, early detection, diagnosis and treatment.

During the period over which this review was written, she was Chief Scientist and Deputy Director of Health Data Research UK, the UK's institute for health data science. She was also Director of the institute's British Heart Foundation Data Science Centre, which aims to improve heart and circulatory health through enabling researchers to access, analyse and derive insights from multiple types and sources of large-scale data. She is Professor of Neurology and Clinical Epidemiology at the University of Edinburgh, where she leads a Scottish population-based longitudinal cohort study (Generation Scotland). In August 2024 she took up a new role as Director of the UK Research and Innovation Adolescent Health Study.

From 2011 to 2019, she led efforts to follow the health of UK Biobank participants through linkage to multiple sources of national health records across England, Scotland and Wales. From 2020, she has worked with NHS Digital (now NHS England) to develop the first trusted research environment to securely hold and enable access for research to linked health data from multiple sources for the whole population of England.

Professor Sudlow is a fellow of the Academy of Medical Sciences and of the Royal Society of Edinburgh. She was awarded an OBE for services to medical research in 2020.

About the support team

A small team of staff at Health Data Research UK (HDR UK) supported the set-up, running and documentation of the multiple stakeholder engagement sessions, online survey and public workshops. These included Lara Edwards, Programme Director for Research Driver Programmes, Dr Lynn Morrice, Operations Director for the BHF Data Science Centre, Dr Ester Bellavia, Patient and Public Involvement and Engagement Manager, Dr Doreen Tembo, Head of Public Involvement and Engagement, and Dr Caroline Bull, Strategy Advisor.

The HDR UK communications and external affairs team provided support for communications with stakeholders, stakeholder briefing sessions to share the main findings and recommendations of the review, liaison with the offices of the review's commissioners, and coordination of copy editing, proof reading and production services.

Udani Samarasekera, freelance health journalist, writer and editor, provided stylistic and copy-editing input.

Several academic, NHS and industry colleagues from across the four nations of the UK, as well as members of HDR UK's senior leadership team,[338] provided feedback on the factual content of the review. However, the lead author takes personal responsibility for the review's content, recommendations and any errors of fact.

338 See https://www.hdruk.ac.uk/about-us/who-we-are/our-senior-leadership-team/.

Appendix 2
Terms of reference for the review

Department of Health & Social Care

Office for National Statistics

NHS

Professor Cathie Sudlow
Chief Scientist
Health Data Research UK

22 March 2023

Dear Cathie

Improving the flow of health data has the potential to significantly improve individual patient outcomes, research for the future and the public's health. This was clear during the COVID-19 pandemic.

There are improvements which can be made to the flow of health data and we would like to invite you to conduct a review which:

- Maps the linkable health data sets across the United Kingdom. This has the support of the Chief Medical Officers from the four nations.
- Outlines any barriers in England and identify solutions to overcome these barriers to sharing data for public benefit, whilst keeping it secure. This has the support of the Secretary of State for Health and Social Care and the Chief Executive of NHS England.

We annex the draft Terms of Reference for the review for you to consider.
This commission has the support of the Department for Health and Social Care, the Office for National Statistics and NHS England among others.
We would be delighted to support you in doing this.

With many thanks for considering this.

Yours sincerely

Professor Chris Whitty **Ian Diamond** **Tim Ferris**

Review of health data
Terms of Reference

1. The Chief Medical Officer for England Professor Chris Whitty, NHS England's National Director of Transformation Dr. Timothy Ferris and the UK's National Statistician Professor Ian Diamond have commissioned Professor Cathie Sudlow to conduct a review of flows of health data. This has the support of the Chief Medical Officers of the four nations of the UK and the Government Chief Scientific Adviser.
2. During COVID-19, we saw the flow of data improve in parts of the health system. Embedding these, other positive changes and enhancing the flow of secure data will help healthcare workers, public health experts, researchers and policy officials to improve patient outcomes.
3. This review should analyse the speed and flow of data within the health system, identify areas where there are barriers and prevent the reversal of progress we have seen to date including what needs to be in place to promote ongoing improvement.
4. Maintaining the principle of patient confidentiality and public confidence is essential.
5. The review should be in two parts:
    - Part A: A mapping of the linkable health data sets across the UK. This includes health data, but also non-health data which has a bearing on health where you think that is relevant.
    - Part B: Outlining any barriers in England, including practical and regulatory, and what we can do to overcome them.
6. The review should cover data related:
    - Direct care.
    - Population health: de-identified data for epidemiology and identified data for case finding.
    - Operational planning: de-identified data with access by NHS and/or government.
    - Research.
7. Reflections on wider data issues which impact on health inequalities would be welcome.
8. It is assumed you will seek Secretariat support from within Health Data Research UK but if this is challenging then we can seek support from within DHSC, NHSE and ONS.
9. You should assemble either as a permanent refence group or consult ad hoc for specific issues whoever you think would be helpful to progress the work.
10. The report should be concluded within six months if possible and be made publicly available. You may prefer to do two reports, with early findings informing action.

Appendix 3
Recent, relevant policy documents, reports or reviews and their areas of coverage

Digital and data capabilities in the health and social care systems

The independent Wachter Review, *Making IT work: harnessing the power of health information technology to improve care in England* (September 2016),[339] focused on the need for adaptive change (including leadership and training) to enable digital maturity to become a reality in the NHS, especially in secondary care. The review made recommendations for a staged – but ultimately comprehensive – roll-out of electronic patient record systems across hospitals.

The National Audit Office report, *Digital transformation in the NHS* (May 2020),[340] provided a thorough and detailed analysis of historical context, plans, spend, proposed budget and progress in digitisation of the NHS in England. It noted the poor previous track record for digital transformation, and significant challenges to be overcome, including outdated, legacy and poorly interoperable systems. It suggested that success would depend on **developing a better understanding of the investment required and a clearer direction for local organisations.**

Laura Wade-Gery's independent report, *Putting data, digital and tech at the heart of transforming the NHS* (November 2021),[341] recommended significant changes in the culture, operating model, skills, capabilities and processes of England's national NHS bodies, aiming to improve citizen and patient outcomes through a more coherent approach to national digital transformation, with appropriate leadership and support from the centre for regional integrated care systems. One consequence has been the inevitably complex, time-consuming and organisationally disruptive merger of NHS England and Improvement, NHS Digital and NHSX.

The independent Goldacre Review, *Better, Broader, Safer: Using Health Data for Research and Analysis* (April 2022),[342] commissioned in February 2021[343] by the then Secretary of State for Health, Matt Hancock, made extensive, detailed recommendations for England on the efficient and safe use of health data for research and analysis to benefit patients and the healthcare sector. Key recommendations included the development and use of a limited number of demonstrably secure trusted research environments for access to NHS data, together with a radical reduction in data dissemination (except where there is explicit consent for this); an open and shared approach to all data curation and analysis code, with accompanying technical documentation, enhancing transparency and efficiency through 'reproducible analytical pipelines'; and the creation of a robust, modern career structure for NHS service analytics, modelled on the government statistical service.

339 https://www.gov.uk/government/publications/using-information-technology-to-improve-the-nhs.
340 https://www.nao.org.uk/reports/the-use-of-digital-technology-in-the-nhs/.
341 https://www.gov.uk/government/publications/putting-data-digital-and-tech-at-the-heart-of-transforming-the-nhs.
342 https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis.
343 https://www.gov.uk/government/news/new-review-into-use-of-health-data-for-research-and-analysis.

The UK Government Department of Health and Social Care (DHSC) policy document, *Data Saves Lives: reshaping health and social care with data* (June 2022),[344] again focused on England, incorporated many recommendations from the Goldacre Review. *Data Saves Lives* highlighted how the use of NHS data at national scale drove the clinical research, health, care, and public health responses to COVID-19. It recognised the need to maintain the momentum of faster, wider, more efficient approaches to data access and use developed during the pandemic, and to apply these more broadly to long-term challenges in health and care. As well as emphasising the ongoing need for digital and technology developments to benefit the care of individual patients, including giving people better access to their own health records, it included commitments to invest in an England-wide network of accredited secure data environments (SDEs, a term interchangeable with TREs, trusted research environments) through NHS England's Data for Research and Development programme.[345] This programme is now building on existing expertise and infrastructure to develop a network of one national and 11 regional secure data environments to safely hold and enable secure, controlled, remote access to health-relevant data for research and innovation.

The UK Government Department of Health and Social Care (DHSC) policy document, *A plan for digital health and social care* (June 2022),[346] published alongside *Data saves lives,* laid out ambitious plans and goals for ongoing and increasing digitisation and digital maturity across the health and social care system in England. Nearer term goals included: all NHS trusts to have an electronic patient record by March 2025; 80% of Care Quality Commission registered social care providers to have digital records by March 2024; a life-long, joined-up health and social care record by March 2025; increasing use of the NHS App as the main interface for interacting with NHS services, with 75% of adults registered with and benefiting from the App's services by March 2024.

The House of Commons Health and Social Care Committee's report on *Digital Transformation in the NHS* (2023),[347] reviewed progress against the UK Government's plans for digital transformation of the health and care service in England. It noted substantial variation in digital capability between health and care organisations and made recommendations on addressing the preponderance of 'legacy' IT in the NHS; the workforce data and digital skills gap; and the challenges of building an inclusive digital health service.

The Scottish and Welsh Governments and Northern Ireland Executive also published **updated, detailed digital and data strategies for health and social care in the devolved administrations (2021-2023)**,[348] in each case laying out their plans for ongoing digital transformation of health and care systems (including plans to create a single national electronic health and care record), as well as to continue to champion the use of data from these systems to drive improvements in care and to support research and innovation.

344 https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data.

345 https://transform.england.nhs.uk/key-tools-and-info/data-saves-lives/accessing-data-for-research-and-analysis/work-in-progress/.

346 https://www.gov.uk/government/publications/a-plan-for-digital-health-and-social-care.

347 https://committees.parliament.uk/publications/40637/documents/198145/default/.

348 https://www.gov.scot/publications/scotlands-digital-health-care-strategy/; https://www.gov.scot/publications/data-strategy-health-social-care-2/; https://www.gov.wales/digital-and-data-strategy-health-and-social-care-wales; https://www.health-ni.gov.uk/publications/digital-strategy-health-and-social-care-northern-ireland-2022-2030.

Strategy for life sciences and clinical research

The UK Government's policy paper on the future of clinical research delivery, *Saving and improving lives: the future of UK clinical research delivery* (March 2021),[349] published on behalf of the DHSC, Scottish and Welsh Governments and the Executive Office of Northern Ireland, outlined a strategy for streamlined and efficient research studies, using innovative, scalable, data-driven and digital approaches to make it straightforward and rewarding for both patients and healthcare staff to participate and contribute.

The UK Government's Life Sciences Vision, *Build back better: our plan for growth* (July 2021),[350] emphasised the importance of building on the UK's globally competitive life sciences and clinical research capabilities combined with its uniquely large and detailed population-based genomic resources and linked health data from the NHS, stating that: *"Over the next decade, high-quality health data will be one of the primary drivers of global Life Sciences research and innovation and improved health outcomes. The NHS has potentially the richest longitudinal health data in the world – but the governance of, and access to, this data must be radically simplified, while simultaneously being made more secure and research-ready, to unlock its full research and innovation potential. We can only achieve this Vision with the full support of patients, the public and NHS, and must build trust into its delivery."*

Lord O'Shaughnessy's independent report, *Commercial clinical trials in the UK* (May 2023),[351] produced on behalf of the UK's Office for Life Sciences (OLS), Department of Science Industry and Technology (DSIT) and the DHSC, showcased examples of UK-led, world-leading, innovative clinical trials delivered through partnerships between the NHS, academia, industry, government and the public, particularly during the pandemic. However, it noted a falling UK performance ranking for the delivery of commercial clinical trials. It identified specific challenges to be solved to change this and made practical recommendations to address them, summarised as *"a public commitment from leaders across the UK demonstrating…our ambition for the NHS to become the world's leading platform for health and life sciences research, followed by a comprehensive plan of reform and a targeted set of key performance indicators against which performance can be judged."* The problems identified included that *"We are failing to take advantage of the NHS's considerable data assets."*

349 https://www.gov.uk/government/publications/the-future-of-uk-clinical-research-delivery/saving-and-improving-lives-the-future-of-uk-clinical-research-delivery.

350 https://www.gov.uk/government/publications/life-sciences-vision.

351 https://www.gov.uk/government/publications/commercial-clinical-trials-in-the-uk-the-lord-oshaughnessy-review/commercial-clinical-trials-in-the-uk-the-lord-oshaughnessy-review-final-report.

The Tony Blair Institute for Global Change independent report, *A new national purpose: harnessing data for health* (May 2024),[352] proposes the creation of a National Data Trust (NDT), a commercial entity, majority owned and controlled by the government and the NHS with investment from industry partners. The NDT would not hold NHS data but would provide a single point of access to a wide range of national and – in due course – regional data assets, together with a range of data concierge, clinical trials and analysis services. The report argues that commercialisation of access to data and services would, in the long term, generate substantial revenue which would be distributed back to the NHS and other organisations providing the data. As a private, rather than public, sector entity, the NDT would have greater flexibility to make long-term investment plans and to attract investment and talented staff. Building and maintaining public trust would be essential. And substantial up-front investment, legislative change and addressing the complexity of the health data ecosystem would all be required.

Data sharing and linkage across government

The UK Government's *National Data Strategy* (2020, intermittently updated)[353] noted that use of data for societal benefit, including improving the efficiency and quality of public services (including healthcare) and boosting science (including life sciences and medical research) is often limited by barriers to access (for example when data are hoarded, when access rights are unclear or when organisations do not make good use of the data they already have). It laid out priority actions to ensure that data from across multiple sectors is fit for purpose, appropriately accessible and used in a safe and trustworthy way by people with the appropriate data skills.

The UK Office for Statistics Regulation report, *Data sharing and linkage for the public good* (July 2023),[354] noted the substantial public benefit potentially achievable through better enabling access to and linkage of data held across government departments. It acknowledged excellent progress in the last several years in creating linked datasets and making them available for research, analysis and statistics, but highlighted ongoing challenges with cross departmental data linkage and access, access to data for researchers outside government, and the need for better understanding of the public's attitude to and confidence in data sharing and linkage.

352 https://institute.global/insights/politics-and-governance/a-new-national-purpose-harnessing-data-for-health.

353 https://www.gov.uk/government/publications/uk-national-data-strategy.

354 https://osr.statisticsauthority.gov.uk/publication/data-sharing-and-linkage-for-the-public-good/.

Appendix 4
Individuals and organisations consulted

This independent review benefited from expert advice and input from the review commissioners and from a reference group, which comprised the Chief Medical Officers of the UK's four nations and senior representation from NHS England, the Office for National Statistics, the Department of Health and Social Care (including the Office for Health Improvement and Disparities) and the UK Health Security Agency.

We are very grateful for the insights, expertise, information, advice, challenge and time provided by several hundred people from a wide range of organisations and groups consulted, or providing input via our online stakeholder survey (see Appendix 5), during this review.

We have listed these organisations and groups (alphabetically) to acknowledge their contributions. Their inclusion here does not imply endorsement of the review's content or recommendations.

10 Downing Street (10DS) Data Science

Academy of Medical Royal Colleges

Administrative Data Research UK

AlbionVC

Albyn Housing Society Ltd

All Wales Therapeutics and Toxicology Centre

Applied Research Collaboration Kent, Surrey and Sussex

Association of British HealthTech Industries

Association of Medical Research Charities

Asthma + Lung UK

Barts Health NHS Trust

Beat Kidney Stones

Bedford Borough Council

Belfast Health & Social Care Trust

BioIndustry Association

Blood Cancer UK

British Geriatrics Society Ageing Data Research Collaborative

British Heart Foundation

British Medical Association

B-Secur

Cancer Research UK

Care Policy and Evaluation Centre, London School of Economics and Political Science

Care Quality Commission

Carnall Farrar

City, University of London

Costello Medical

Cystic Fibrosis Trust

Data and Analytics Research Environments UK

DATA-CAN

Dementias Platform UK

Department for Science, Innovation and Technology

Department of Health and Social Care (including Data Policy Directorate, DHSC/NHSE Digital Policy Unit, National Screening Programmes and Office for Health Improvement and Disparities)

Department of Health Northern Ireland

Diabetes UK

- Economic and Social Research Council, UK Research and Innovation
- Faculty of Medicine, Imperial College London
- FITFILE
- Frimley Integrated Health
- Generation Scotland
- Genomics England
- GlaxoSmithKline
- Great Ormond Street Institute of Child Health
- Health and Social Care Data Institute Northern Ireland
- Health and Social Care Northern Ireland, including Honest Broker Service
- Health Data Research UK (including Board of Trustees, Public Advisory Group, and Strategy and Integration Group)
- Health Innovation Research Alliance Northern Ireland
- Health Research Authority
- Hereford and Worcester Integrated Care Board
- Imperial College London
- Innovative Healthcare Delivery Programme Scotland
- Intensive Care National Audit and Research Centre
- IQVIA
- King's College London
- Lancaster University (including Lancaster Medical School)
- Lane, Clark and Peacock
- LifeArc
- Lifebit
- London School of Hygiene and Tropical Medicine
- London Strategic Information Governance Networks' Forum
- Manchester University NHS Foundation Trust
- medConfidential
- Medical Research Council, UK Research and Innovation
- Medicines and Healthcare products Regulatory Agency (including Clinical Practice Research Datalink Data Services)
- Mendelian Ltd
- Midlands Partnership University NHS Foundation Trust
- Mydex Community Interest Company
- National Cancer Audit Collaborating Centre
- National Centre for Social Research
- National Data Guardian Office
- National Institute for Health and Care Excellence
- National Institute for Health and Care Research
- NEC Software Solutions UK
- NHS Bedfordshire
- NHS Business Services Authority
- NHS England (including Data and Analytics Sub-directorate, Data Enabled Research Advisory Group, Data for Research and Development Programme, National Disease Registration Service, Pathology and Laboratory Medicine Informatics, and Transformation Directorate)

NHS England London Region

NHS Lothian

NHS South East London Integrated Care Board, with input from Lambeth Council Public Health Directorate

NHS Wales (including National group for Digital and Data Knowledge)

NHS West Yorkshire Integrated Care Board

NHSE Regional London Secure Data Environment

NI Cancer Research Consumer Forum

NIHR Clinical Research Network North West Coast

NIHR Clinical Research Network Wessex

North West Region Secure Data Environment Programme

Northern Ireland Statistics and Research Agency

Novartis

Office for Life Sciences

Office for National Statistics

Office for Statistics Regulation

OpenSAFELY/Bennett Institute for Applied Data Science

Optimum Patient Care

Ordnance Survey

Our Future Health

Palantir

Population Health Partners

Public Health Scotland

Public Health Wales

Queen Mary University of London

Research Data Scotland

RISG Consulting

Royal Brompton Hospital

Royal College of General Practitioners (RCGP) (including RCGP Health Informatics Group and Oxford RCGP Research and Surveillance Centre)

Royal College of Radiologists

Royal College of Surgeons of England (including Clinical Effectiveness Unit)

SAIL Databank

Salford Royal NHS Foundation Trust (Northern Care Alliance)

SCONe (Scottish Collaborative Optometry-Ophthalmology Network e-research)

Scottish Government, including Chief Medical Officer Directorate and Chief Scientist Office

Swansea University

The Association of British Insurers

The Association of the British Pharmaceutical Industry

The Fatherhood Institute

The Fragile X Society

The Health Foundation

The Institution of Engineering and Technology

Tony Blair Institute for Global Change

Triscribe Ltd

UK Biobank

UK Health Data Research Alliance (including Pan-UK Data Governance Steering Group)

UK Health Security Agency

UK Longitudinal Linkage Collaboration

UK National Screening Committee

UK Renal Health Data Network

UK Renal Registry

UK Research and Innovation

UK Statistics Authority

Understanding Patient Data

University College London (including MRC Clinical Trials Unit: Institute of Clinical Trials and Methodology and Centre for Longitudinal Study Information and User Support (CeLSIUS))

University College London Hospitals NHS Foundation Trust

University Hospital of Wales, Cardiff

University Hospitals Birmingham, University of Birmingham, UK Organ Donation & Transplantation Research Network

University Hospitals of Derby and Burton NHS Foundation Trust

University of Aberdeen

University of Bristol

University of Cambridge

University of Dundee

University of Edinburgh (including Centre for Clinical Brain Sciences, DataLoch: NHS Lothian Data Safehaven, Scottish Centre for Administrative Data Research, and Usher Institute)

University of Essex

University of Glasgow

University of Hull

University of Leeds (including Consumer Data Research Centre)

University of Liverpool (including DynAIRx: Artificial Intelligence for dynamic prescribing optimisation and care integration in multimorbidity)

University of Manchester

University of Northumbria

University of Nottingham

University of Oxford

University of Reading

University of Strathclyde

University of Ulster

University of Warwick

University of Westminster

University of Winchester

University of York (including York Trials Unit)

use MY data

Wellcome Trust

Welsh Government

Yorkshire Cancer Research

Appendix 5
Findings from online survey and public workshops

This appendix summarises key findings from the online survey (section 1) and public-facing workshops (section 2) contributing to the review.

Section 1: Online survey

An online survey, accessible through the Health Data Research UK (HDR UK) website, was conducted from 30 May to 7 July 2023 to capture perspectives on the use of health data from across the UK. The survey aimed to:

- Assess current views on the use of health data in research and healthcare.
- Identify priorities for data availability to support advances in these fields.
- Determine barriers and propose solutions for maximising the potential of health data in the UK.
- Gather additional relevant information for the review.

A total of 178 responses were received. Some were from individuals, while others were collective submissions by teams or organisations. These responses were categorised as follows: academia (n=80), charity organisations (n=9), funders (n=2), government (n=10), industry (n=20), National Health Service (NHS) (n=33), individual members of the public (n=22), and organisations representing the interests of patients and the public (n=2).

Key findings

Views on health data

- Survey respondents agreed that health data represents a valuable yet under-utilised resource. Its use in real time was seen as crucial for providing insights to tackle disease outbreaks, improve healthcare delivery, and facilitate resource management in the NHS.
- Significant issues raised included the potential for data exploitation by pharmaceutical companies and insurance providers, challenges surrounding data representativeness, accessibility and fragmentation, data quality, and public trust in data security.

Dataset priorities

- Among all responses received, primary care data were widely acknowledged for their comprehensive and longitudinal nature. Accessing these data were regarded as pivotal in advancing preventive healthcare initiatives and supporting research. There were concerns regarding data quality, accessibility and effective management, and the need for robust de-identification processes.
- Survey respondents emphasised the importance of integrating NHS and non-NHS administrative data to drive research and healthcare initiatives forward. This integration not only holds the potential to uncover inequalities but also to tailor services to the needs of a diverse population.

Overcoming barriers

- Survey respondents highlighted the critical role of data regulations in shaping access to health data. Current laws and complexities in information governance are often cited as barriers. Respondents called for standardised agreements, streamlined procedures and legislative frameworks that facilitate prompt and secure data sharing.
- Several responses highlighted the issues of data ownership and control, citing ambiguities surrounding NHS data governance as a significant barrier to effective data sharing. There was a consensus in favour of establishing a central regulatory body responsible for overseeing data access, ensuring standardisation, and enforcing compliance.
- Overcoming siloed mentalities and promoting a collaborative data ecosystem were deemed necessary to fully unlock the potential of health data. Providing clear incentives and recognising the NHS as a research organisation were identified as potential solutions.
- Survey respondents identified the need for better infrastructure as a key issue, emphasising the need for improved interoperability, standardisation, and long-term funding. Many also proposed centralised data hubs to streamline data access and management.
- Respondents emphasised the importance of training and support in ensuring data quality. They advocated for comprehensive training programmes in data governance and management for healthcare professionals and researchers. Additionally, creating clear and attractive career pathways in the health data field was seen as vital for building capacity and expertise.
- Public trust emerged as a recurring theme. Survey respondents stressed the importance of transparent communication about data security and the benefits of data sharing. Engaging with underserved communities and addressing privacy concerns through consistent dialogue and public engagement were highlighted as key strategies.

Section 2: Workshops with the public

A total of 100 members of the public participated in two online workshops conducted on 23 August (n=71) and 6 September (n=29), 2023. Participants were from a wide range of age groups spanning 16 to >65 years old; almost two thirds were female; just under two thirds were White and the remainder were from a range of other ethnic groups (including Black African, Asian Pakistani, Asian Bangladeshi, Asian Indian, Arabic, Black Caribbean, Chinese, Mixed and other ethnicities); just over 50% reported some form of disability; around one third were retired, 40% employed or self-employed, 10% unable to work, 6% unemployed, and the remainder were students, interns, informal carers or working unpaid at home.

The workshops were structured to facilitate in-depth discussions on:

- the advantages and concerns associated with the use of data for the public's benefit;
- priority data needs for health and care planning, public and population health management, and research;
- identifying obstacles and potential solutions to better use of data for patient and public benefit.

The following sections outline key insights from the analysis of verbatim transcripts from group discussions recorded with attendees' consent.

Key findings

Inclusive and innovative public engagement initiatives

- Participants underscored the importance of clear and accessible communication to highlight the benefits of data sharing.
- They suggested that storytelling and diverse communication strategies, including social media, documentaries, and public advertisements, can effectively reach a wide audience.
- Ensuring equity and inclusivity in communication was perceived crucial, especially for underrepresented communities.
- Creating clear and simple, plain language guidelines on health data was considered essential to enhance health data literacy among the public.

Data ownership and accuracy: barriers and ways forward

- Workshop participants emphasised the importance of patients having ownership and control over their health data, including the ability to rectify inaccuracies if necessary.
- Challenges related to accessing personal health data were mainly attributed to technological limitations within the NHS and the reluctance of general practitioners (GPs) to grant access.
- There was a consensus on the need for accurate and comprehensive data collection from across all parts of the healthcare system, not just general practice.
- Potential solutions include providing adequate resources and incentives for GPs and ensuring health data training opportunities are available to a wide range of healthcare professionals.

Prioritising datasets for linkage

- Workshop attendees prioritised access to general practice data and emphasised the need for linking it with other data sources, such as secondary and social care data.
- They also considered obtaining and linking data from care homes, home care, social workers, and mental healthcare settings crucial for enhancing the quality of care, especially for those in long-term and mental healthcare settings.
- Administrative data, including socio-economic and environmental information, was perceived as necessary for facilitating more people-centric service delivery.
- Concerns raised focused on the inclusiveness and representativeness of current datasets, with some attendees cautioning against linking existing datasets without addressing these underlying issues.

Obstacles and potential solutions to better use of data for patients benefit

- Workshop participants felt that the lack of integration of different sources of health data posed a significant obstacle to a comprehensive understanding of patients' medical histories.
- As a result, standardising and improving healthcare systems to facilitate data sharing was deemed necessary, with financial investments highlighted as a key requirement.
- Concerns were raised regarding trust in private healthcare providers and commercial organisations using patient data, with calls for robust controls to safeguard privacy and ensure ethical use of data.
- Clear and transparent information for members of the public on data anonymisation and data flow processes was also recommended, to foster better understanding and awareness.

Conclusions

In conclusion, while health data holds significant promise for advancing research and healthcare, realising its full potential requires regulatory, governance, infrastructural, and trust-related barriers to be addressed, as highlighted by the survey responses. Collaborative efforts and a commitment to transparency and public engagement are deemed essential to overcoming these challenges. Additionally, findings from the workshops provide insights into the public's perspective on the use of data in research and healthcare, underscoring the need for improved communication, data accuracy, inclusiveness, and system integration to better serve patients and the public.

Appendix 6
Examples of the UK's many prospective longitudinal cohorts

Cohort	Number of participants	Data and samples collected	Linked data from NHS and other administrative sources	Researcher access
VERY LARGE POPULATION-BASED RESEARCH RESOURCES				
UK Biobank www.ukbiobank.ac.uk	~500,000 aged 40-69 years at recruitment 2006-2010.	Baseline questionnaire, physical measures and bio-samples (blood, urine, saliva). Additional data via online questionnaires, wrist worn accelerometry, brain and body imaging from large subsets. Genotyping, genetic sequencing, and biochemistry assays for all participants; proteomic and metabolomic data for large subsets.	Regularly updated hospital episodes, death register, cancer register data from England, Scotland and Wales. General practice data, specialist audit and other health data being sought. Other administrative data will be sought in future.	Streamlined process for access to de-identified data for research in the public interest by registered bona fide researchers worldwide. Data increasingly accessed via UK Biobank's secure Research Access Platform.
Our Future Health ourfuturehealth.org.uk	Target of 5 million adults; over 1.5 million recruited by June 2024.	Baseline questionnaire, physical measures and blood sample collection.	Linked data (as for UK Biobank) being sought.	Access to de-identified data for registered researchers worldwide via Our Future Health's Trusted Research Environment for basic science, epidemiological discovery and aetiological research. Opportunities in the pipeline for translational research with re-contactable participants.
Genomics England initiatives, including 100,000 Genomes Project and Newborn Genomes Programme www.genomicsengland.co.uk	85,000 NHS patients with rare disease or cancer in 100,000 Genomes Project. Target of 100,000 newborn babies in the Newborn Genomes Programme.	Extensive clinical and genome sequencing data.	Linked data (as for UK Biobank).	Access to de-identified data for approved researchers worldwide within the Genomics England secure research environment.

Cohort	Number of participants	Data and samples collected	Linked data from NHS and other administrative sources	Researcher access
SMALLER POPULATION-BASED COHORTS				
ALSPAC[355] www.bristol.ac.uk/alspac	Multigenerational study of 14,500 people born in the former county of Avon in 1991-1992 as well as their parents and children.	Participant questionnaires, physical measures, bio-samples (blood, urine, hair, nails, saliva) and results of sample assays (genotyping, genomic sequencing, gene expression, methylation).	Wide range of linked health and other administrative data.	Access to de-identified data for approved researchers via ALSPAC access process or via the Longitudinal Linkage Collaboration's trusted research environment.
Generation Scotland www.genscot.ed.ac.uk	24,000 adults recruited from 7000 families 2006-2010. Recruitment of a further 20,000 participants age 12+ underway (12,000 by June 2024).	Participant questionnaires, physical measures and blood samples on original recruits. Online questionnaires and saliva on newer participants. More detailed phenotype data including imaging on subsets. Genotyping, DNA methylation and proteomics on all/ most participants.	Wide range of linked health data and linkage to other administrative records in the pipeline.	Access to de-identified data for approved researchers via Generation Scotland access process, via Dementias Platform UK[356] or via the Longitudinal Linkage Collaboration's SDE.[357]
DISEASE- OR DOMAIN-SPECIFIC COHORTS				
UK HFpEF[358] **Registry** www.ukhfpef.org	>600 adults with HFpEF from centres across the UK in pilot phase, now aiming to expand to 7000 patients.	Detailed clinical information, including cardiac imaging, and bio-samples for future assays.	Linkage to a wide range of national health data sources planned.	Access to de-identified data for approved researchers will be via the BHF Data Science Centre's Cardiometabolic Cohorts SDE.[359]
BSRBR-RA[360] www.bsrbr.org	>30,000 patients, recruited from across the UK since 2001, with rheumatoid arthritis and prescribed targeted immune therapy.	Detailed clinical information from rheumatology teams and patient questionnaires.	Linkage to national hospital episodes, cancer registry and death register data across the four nations of the UK.	Access to de-identified data via the British Society for Rheumatology.

355 ALSPAC: Avon Longitudinal Study of Parents and Children.

356 See https://www.dementiasplatform.uk/data-portal.

357 See https://ukllc.ac.uk/.

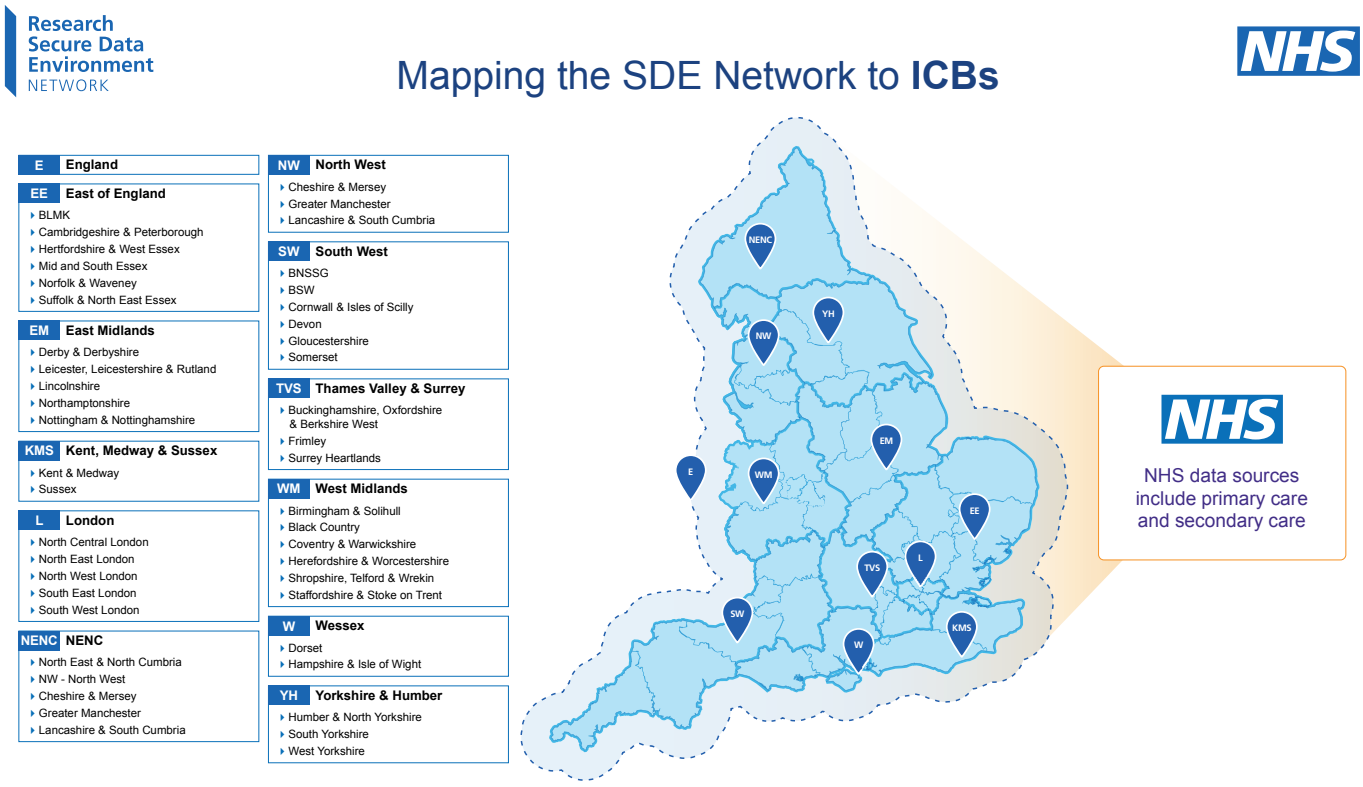
358 HFpEF: heart failure with preserved ejection fraction (a poorly understood type of heart failure affecting about 50% of patients with heart failure).

359 See https://bhfdatasciencecentre.org/areas/cohorts/.

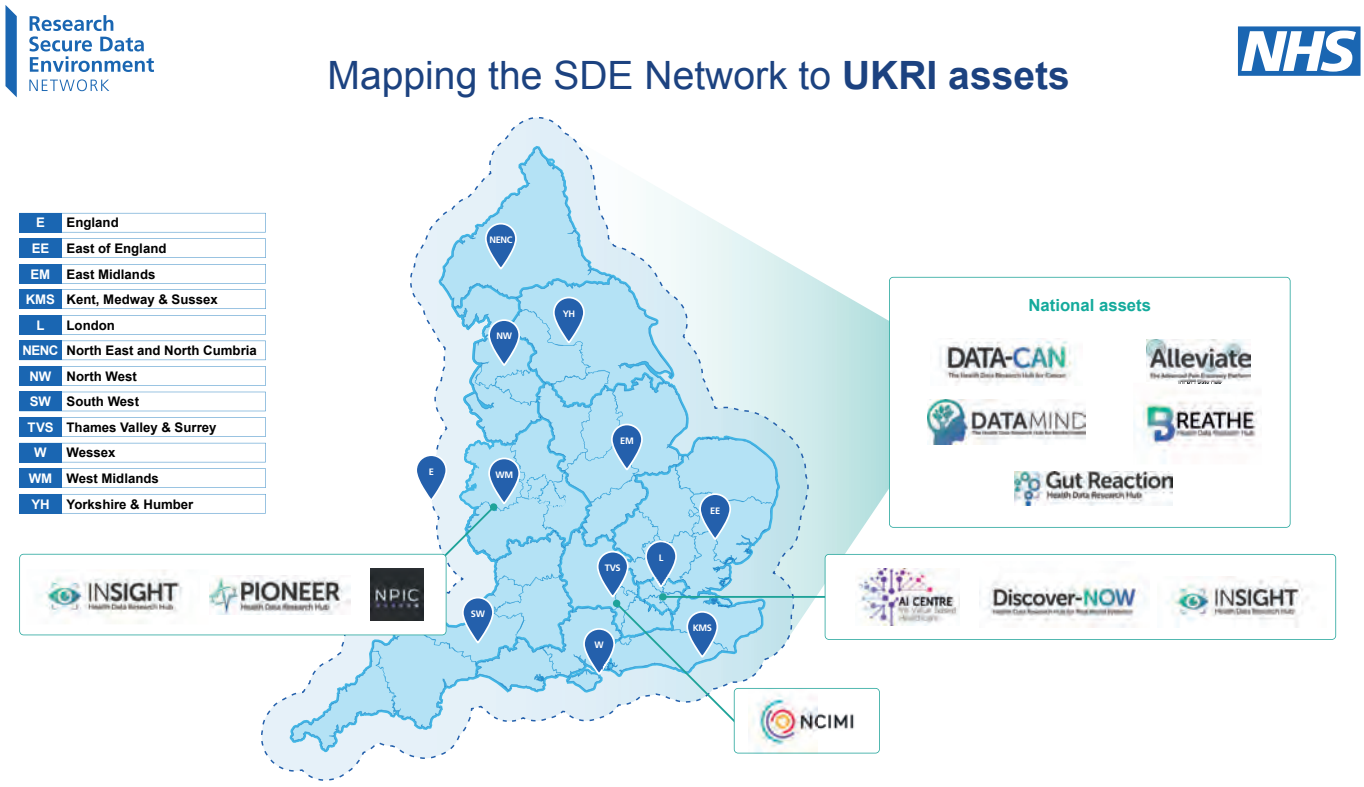
360 British Society of Rheumatology Biologics Register- Rheumatoid Arthritis.

Appendix 7

Mapping England's NHS regional secure data environment network to existing UK research infrastructure



ICBs: integrated care boards



UKRI: UK Research and Innovation

DATA-CAN, BREATHE; Gut Reaction, INSIGHT, PIONEER and Discover-NOW are health data research hubs, supported from 2019 to 2023 by the UKRI Industrial Strategy Challenge Fund; DATAMIND and Alleviate are Medical Research Council-funded health data research hubs. All the health data research hubs are delivered in partnership with Health Data Research UK.

NPIC: National Pathology Imaging Co-operative

NCIMI: National Consortium of Intelligent Medical imaging

AI CENTRE: AI Centre for value-based healthcare

Research Secure Data Environment NETWORK

Mapping the SDE Network to **NIHR assets**

NHS

E England

EE East of England
- East of England RDN
- Cambridge BRC

EM East Midlands
- East Midlands RDN
- Nottingham BRC
- Leicester BRC

KMS Kent, Medway & Sussex
- South East RDN

L London
- North London RDN
- South London RDN
- Barts BRC
- Great Ormond Street Hospital BRC
- Imperial BRC
- Maudsley BRC
- Moorfields BRC
- The Royal Marsden BRC
- University College London Hospitals BRC

NENC NENC
- North East and North Cumbria RDN
- Newcastle BRC

NW North West
- North West RRDN
- Manchester BRC

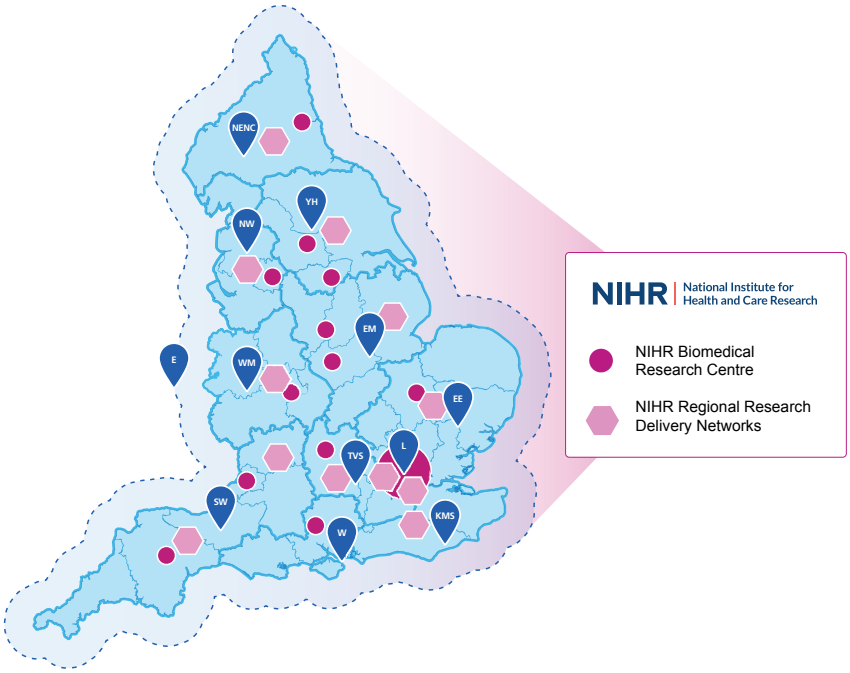
SW South West
- South West Peninsula RDN
- South West Central RDN
- Bristol BRC
- Exeter BRC

TVS Thames Valley & Surrey
- South Central RDN
- Oxford BRC
- Oxford Health BRC

WM West Midlands
- West Midlands RDN
- Birmingham BRC

W Wessex
- South Central RDN
- Southampton BRC

YH Yorkshire & Humber
- Yorkshire and Humber RDN
- Leeds BRC
- Sheffield BRC

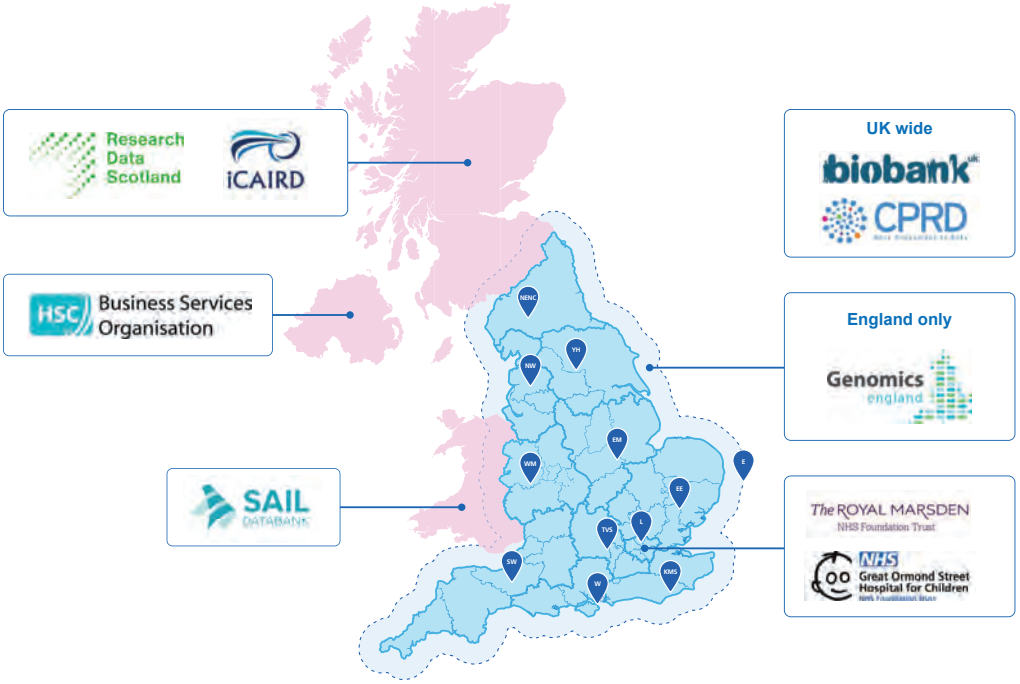


Research Secure Data Environment NETWORK

The SDE Network and **existing ecosystem**

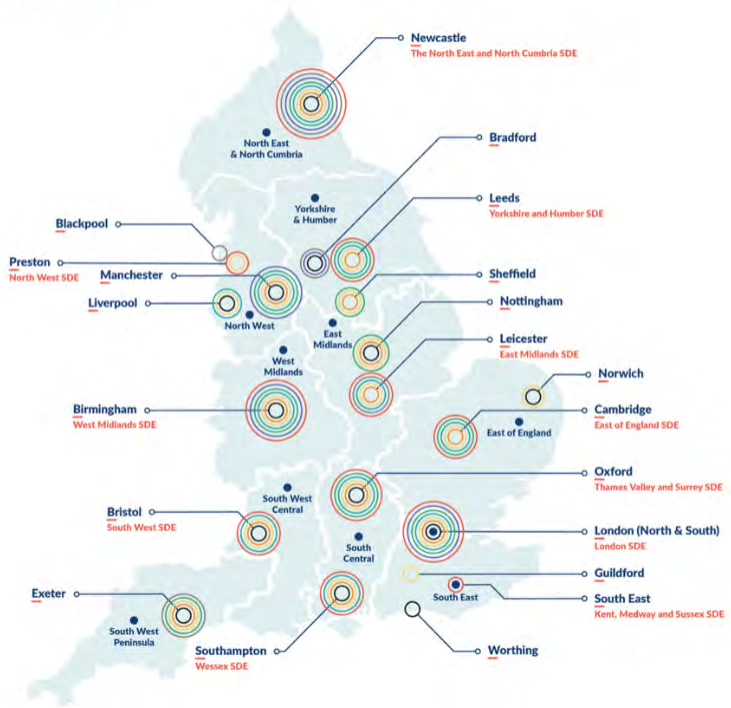
NHS

E	England
EE	East of England
EM	East Midlands
KMS	Kent, Medway & Sussex
L	London
NENC	North East and North Cumbria
NW	North West
SW	South West
TVS	Thames Valley & Surrey
W	Wessex
WM	West Midlands
YH	Yorkshire & Humber



ICAIRD: Industrial Centre for Artificial Intelligence Research in Digital Diagnostics (Scotland); CPRD: Clinical Practice Research Datalink; HSC: Department of Health and Social Care (Northern Ireland); SAIL: Secure Anonymised Information Linkage (Wales)

NIHR INFRASTRUCTURE AND NHS SDE NETWORK FROM OCTOBER 2024



KEY

- Regional Research Delivery Network
- Applied Research Collaborations
- Biomedical Research Centres
- Clinical Research Facilities
- Experimental Cancer Medicine Centres
- HealthTech Research Centres
- Patient Safety Research Collaborations
- Patient Recruitment Centres
- NHS Regional Secure Data Environment (SDE)

Appendix 8
Linked health data resources with English general practice data as a core component

Resource	Brief description	Population coverage	Detail of primary care data available	Additional linked data available	Application for access	Mechanism of data access to data and reproducible code	Additional services	Other comments
NHS England (NHSE) GP Data for Pandemic Planning and Research (GDPPR).[361]	Established 2020, initially to help meet urgent data needs of the COVID-19 pandemic. Data extracted from general practice computer systems into NHSE secure systems by the General Practice Extraction Service.	Near complete – >98% of English practices (all English general practice computer system suppliers).	Large subset of all coded data items. Initially updated fortnightly, moved to monthly in 2024. Only available for COVID-19-related analyses at present.	Linkage of other health datasets available via NHSE Data Access Request Service (DARS) includes hospital episodes, registered deaths, community dispensed medicines, COVID-19 test and vaccination data, hospital e-prescribing, maternity and mental health services data, several specialist audits/registries.	Via NHSE DARS. Consortium model operated by BHF Data Science Centre (DSC)[362] currently enables rapid access for many projects that would otherwise have very long waits.	Most access now within NHSE SDE. Some secure, approved, external data transfer has occurred.	NHSE provides a range of other data services (e.g., NHS Digitrials) but these do not currently use GDPPR data.	Plans stalled in 2021 to replace all General Practice Extraction Service extracts with an efficient, single, comprehensive extraction system to support a wide range of uses for care and research across many health conditions beyond COVID-19 (see section 5.3).
OpenSAFELY.[363]	New platform established 2020, initially to help meet urgent data needs of the COVID-19 pandemic. Initiated as a collaboration between the Bennet Institute (University of Oxford), London School of Hygiene (LSHTM), NHSX, and TPP, with later addition of EMIS, working on behalf of NHS England.	Near complete – all TPP and EMIS practices (although access to data within EMIS systems temporarily suspended in 2024).	Comprehensive coded data within TPP (and EMIS) systems. Near-real-time data. Only available for COVID-19-related analyses at present.	Linkage of other health datasets includes hospital episodes, registered deaths, COVID-19 tests, ISARIC cohort study, UK Renal Registry. Able to map and link new datasets as required.	Access requests to team@opensafely.org for the pilot programme of external users. Analysts require advanced computational skills but receive support through co-piloting service.	Analysts do not access person-level data. Instead, analysis code developed against synthetic data and then run against de-identified near-real-time data within TPP/EMIS systems.		Plans to extend service for uses beyond those related to COVID-19 (see section 5.1).
Clinical Practice Research Datalink (CPRD).[364]	Long-established real-world data service that collects and links anonymised patient data from a network of general practices to support public health and clinical studies. Delivered by the Medicines and Healthcare products Regulatory Agency (MHRA) with support from the National Institute for Health Research (NIHR).	Partial – about 30% of all UK general practices (currently InPractice Systems Vision and EMIS practices).	Comprehensive coded data from participating practices. Data regularly updated.	Linkage of other health datasets includes hospital episodes, registered deaths, cancer registry data, diagnostic imaging dataset, mental health services dataset, COVID-19 test data.	Employee of a CPRD approved client organisation can apply for access to data via CPRD Research Data Governance process.	Following 35 years of providing proven safe, secure access via secure transfer, now developed SDE (CPRD Safe) for data access.	Wide range of data services include data quality and verification and support for clinical trials through its general practice network.	Extensively used by MHRA, commercial and academic users. Self-sustaining through a cost-recovery pricing model.

361 See https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research.

362 See https://bhfdatasciencecentre.org/areas/cvd-covid-uk-covid-impact/.

363 See https://opensafely.org/.

364 See https://www.cprd.com/.

Resource	Brief description	Population coverage	Detail of primary care data available	Additional linked data available	Application for access	Mechanism of data access to data and reproducible code	Additional services	Other comments
Oxford RCGP Research and Surveillance Centre.[365]	Long established source of information, analysis and interpretation of primary care data. Collects data from member general practices in England to create an accessible repository of data for health research. Data hosted at University of Oxford.	Partial – covering around 30% of people registered with a participating practice using the InPractice Systems Vision, TPP or EMIS systems.	Coded data extracted from general practice systems.	Linkage of other health datasets on a per project basis.	Researchers interested in conducting primary care/linked data studies must apply for approval from the Primary Care Hosted Research Datasets Independent Scientific Committee.	Secure, remote access to de-identified data within University of Oxford-hosted ORCHID-E secure environment planned (currently under development).	Infectious disease surveillance services (including bio-samples) via participating practices. Clinical trial remote follow-up services currently in development.	Surveillance services have supported UK Health Security Agency.
QResearch.[366]	Long established, ethically approved database derived from the anonymised health records of EMIS general practices. Data held and accessed securely on servers at the University of Oxford.	Partial – large subset of EMIS general practices.	Comprehensive coded data from EMIS system for contributing practices. Data regularly updated.	Linkage of other health datasets includes hospital episodes, registered deaths, registered cancers, ICNARC critical care data, COVID-19 test and vaccination data, and pregnancy registry.	Access to academic researchers, subject to ethical committee and QResearch Scientific Committee approval. Researchers may need to seek approval for additional linked data.	Remote access to de-identified data within Oxford-hosted Q Research environment.		

365 See https://orchid.phc.ox.ac.uk/.

366 See https://www.qresearch.org/.

Appendix 9
Priority system requirements (focusing on England)

System need	What is needed and why?	What are the main barriers?	How can they be overcome?
Increase speed, timeliness and scope of data access.	• Relevant organisations acknowledge that ecosystem complexity and fragmentation is a problem and must be reduced. • Coordinated joint strategy for health data as critical national infrastructure. • Long-term planning and investment since multiple complex systems and processes cannot simply be replaced and require more than a quick fix. • Single, streamlined, national health data access system • To realise the multiple benefits of access to and use of health data for healthcare, public health, life sciences and wider society.	• Complexity and fragmentation (organisational, computer system, transactional, legal and regulatory). • Insufficient political and organisational incentives for long-term planning and investment. • Capacity gaps – especially information governance (IG) specialists and data management/curation experts – at least partly due to difficulty attracting appropriately qualified specialists and reduced headcount with loss of specialist staff following NHSE merger with NHSD, NHSX and HEE. • Limited funding/resources.	• To develop coordinated joint heath data infrastructure strategy, major national organisations[367] should • Take joint responsibility and accountability for reducing ecosystem complexity and fragmentation. • Appoint senior executive leader, reporting directly to CEOs of NHSE and NIHR, with responsibility and ring-fenced budget for a national health data service.[368] • Review and modify commissioning, contracts, and technical processes to reduce NHS organisational, computer system and transactional complexity (NHSE and DHSC). • Resist temptation to establish new initiatives that may worsen rather than reduce ecosystem complexity. • Consider where new/revised legislation will genuinely help but avoid it where clear, consistent interpretation of current legal and regulatory processes would be as effective. • Streamline and standardise data governance and access, including through introducing a single national health data access system, with performance monitoring, targets and incentives that maximise beneficial uses of data.[369] • Address capacity gaps, especially within NHSE. - rebuild and enhance capacity in specialist health data IG and in data management and curation in national, regional and local NHS settings. - strengthen and formalise partnership between NHSE and health data research community to enable innovative, streamlined, user-informed system design and NHSE secondments to help address headcount constraints. • Long-term planning and investment–non-partisan publicly funded long-term health data infrastructure investment not solely driven by crisis management or unrealistic expectations on delivery timelines. • Practical and acceptable data infrastructure investment and cost models. - strategic precompetitive industry investment models.[370] - transparent costs for public sector, non-profit and for-profit uses.

367 These include NHS England (NHSE), the Department of Health and Social Care (DHSC), the Department of Science Industry and Technology (DSIT), UK Research and Innovation (UKRI), the National Institute for Health Research (NIHR), the Office for Life Sciences (OLS), the Association of Medical Research Charities (AMRC), national research institutes/organisations such as Health Data Research UK (HDRUK) and Administrative Data Research UK (ADRUK), the UK Health Security Agency (UKHSA), Office for National Statistics (ONS), Health Research Authority (HRA), Medicines and Healthcare products Regulatory Agency (MHRA) and the National Institute for Health and Care Excellence (NICE).

368 Essential qualities include ability to lead cultural change; and credibility among clinical, policy, research and data science/technical expert communities.

369 This could be modelled on the Integrated Research Application System, which has had a major impact on the efficiency, speed and transparency of the research approval process. IRAS is a single application system for permissions and approvals for health and care research in the UK, avoiding duplication, providing guidance and advice, and capturing the information required for approvals from multiple bodies (e.g. Confidentiality Advisory Group (CAG), HRA, MHRA, NHS / HSC R&D offices and Research Ethics Committees). IRAS is updated regularly in response to applicant feedback and sets (and reports on) acceptably rapid application turnaround targets.

370 E.g., UK Biobank and Our Future Health have used such models successfully.

System need	What is needed and why?	What are the main barriers?	How can they be overcome?
Maintain broader access achieved during pandemic.	• UK response to the pandemic would have been better informed by robust data if better flows, linkage and access to health data had been in place prior to the pandemic. • Improved flows, linkage and access to health data during the COVID-19 pandemic brought gains in efficiency, productivity and positive impact on patient care and health policy. • Improvements should be maintained and enhanced to better inform current healthcare and public health policy as well as to prepare for future pandemics and other health shocks.	• Drift back to pre-pandemic behaviours: less co-operation, collaboration and 'can-do' across relevant national organisations. • Loss of clear alignment of incentives and common goals, with reduced common sense of urgency about today's challenges vs those faced during the pandemic. • Effects of recent NHS organisational change and upcoming election on political and health system activity and productivity. • Limited funding/resources.	• Multi-pronged, multi-organisational ongoing communication strategy to better articulate for multiple audiences how data can help solve non-COVID-related healthcare and public health challenges.[371] • Extend legal gateways successfully applied during COVID to other health-related uses of data.[372] • Recognise and apply lessons learned during the pandemic rather than simply returning to pre-pandemic business as usual. E.g., we saw that: - multiple national organisations can align and work together towards common goals, with faster, more efficient, higher productivity. - faster, broader data access for research and analysis during pandemic delivered successfully without data security or privacy breaches.
Maintain and enhance national data assets.	• Major public benefit and UK life science national competitiveness depend on our national scale, linkable health data. • Prioritise key generic data from different sources (especially general practice, hospital, medicines and mortality data) as a foundation onto which specialist data can be layered.	• Investment in granular local and regional data infrastructure welcome but must complement and not aim to replace or distract from important national data capabilities. • Lack of support from regions, due to past and ongoing difficulties accessing national data for regional planning and research. • Limited funding/resources.	• Clear roadmap for development and – where appropriate – integration of new national NHSE data infrastructures. • Set priorities for incorporating and enabling access to national generic and domain-specific data assets. • Ensure regions can rapidly access and benefit from national data assets relevant to regional planning, research and innovation. • Consistent and logical development of regional data infrastructures that build on national data assets, adding complementary data detail and granularity not available at national scale. • Contain/reduce costs by avoiding duplication across national and regional infrastructures.
Maintain and improve capability for secure data transfer.	• While unnecessary data travel should be minimised, transfer of data between secure locations is needed for some highly beneficial data uses. • This includes secure transfer of linked health data from national systems for integration into databases of 'consented' cohorts, clinical trials and other research studies.	• Insufficient recognition across multiple relevant organisations of ongoing need for data transfer mechanisms. • Lack of proper integration of requirements for secure data transfer between secure locations into national health data infrastructure resourcing and planning efforts.	• Engage with relevant public and private research trials, cohorts and clinical studies organisations and communities to understand needs. • Include external data transfer mechanism as part of single national data access and provisioning system.

371 Including: 'pandemics' of cancer, cardiovascular disease, diabetes, obesity, mental health conditions and many others; NHS waiting times, workforce crisis and winter pressures.

372 E.g. Secretary of State directions and COPI notices.

System need	What is needed and why?	What are the main barriers?	How can they be overcome?
Improve data usability (through better data quality, metadata, standardisation and linkage).	• Improved data usability across the domains of quality, metadata, standardisation, reproducibility and linkage will improve efficiency, accuracy and relevance of insights and help to identify and reduce inequalities in the data.	• Insufficient availability of metadata.[373] • Suboptimal data quality. • Insufficient standardisation of data collection, formats and terminologies. • Need for improved reproducibility. • Approaches used by national custodians for linking data from different sources may be poorly described or opaque. • Linkage at place (UPRN) as well as person level often unavailable but needed for many beneficial uses. • Data on accuracy of matching at person or record level rarely provided, limiting the validity and reliability of the insights from these data.	• Improved metadata and organisational accountability for keeping it accurate and up to date. • Data quality improvement strategies. - ensure wider data use as this improves quality. - improve digital maturity of health and social care systems. - training and quality assurance to improve quality of data generation. - accelerate efforts for patients to access, question and correct their own health records. • Standardise data collection, formats and terminologies. • Embed reproducible data curation and analysis pipelines as standard within SDEs. Sharing of code within and between SDEs should be a condition for data access. • Enhance data linkage capabilities through. - transparent descriptions of data linkage methods. - UPRN as well as person-level linkage needed to study effects on health of location-associated social and environmental exposures and risk factors. - person- and record-level indicators of accuracy of matching to national personal demographic data needed to better define population denominator in multiple studies, and to better understand and improve linkage quality by age, sex, geography, ethnicity and deprivation. • Improve user experience and reduce costs through. - providing data curation support. - mandating reproducibility and sharing of protocols, code and algorithms. - 'green coding' practices to increase efficient use of shared compute.
Make secure data environments the most attractive option for most uses and users.	• If access to data within SDEs is to become the default for most health data access then they need to be the most attractive option for most uses and users. This is not currently the case.	• SDEs vary, including in maturity, user and data security and authentication protocols, hardware and software provided, access to memory and compute, data management protocols, user support and costs. • Many researchers perceive accessing and analysing data in SDEs to be more difficult than on organisational servers or personal computer.	• National SDE accreditation: should build on established and widely supported schemes, in particular UK Statistics Authority.[374] • National SDE standards: e.g. SATRE.[375] • Need policy to avoid too many SDEs. • Agile adaptation to user feedback and user needs. • Promote and incentivise positive user behaviours (e.g. efficient coding, sharing of protocols, code and algorithms, 'green' coding), so that all benefit.

373 Metadata is information about the data, including dictionaries and other characteristics of data, such as coverage and missingness.

374 See https://uksa.statisticsauthority.gov.uk/digitaleconomyact-research-statistics/better-access-to-data-for-research-information-for-processors/.

375 Standard Architecture for Trusted Research Environments (https://satre-specification.readthedocs.io/en/stable/).

System need	What is needed and why?	What are the main barriers?	How can they be overcome?
Improve transparency for and meaningful engagement with patients, public and healthcare professionals, policymakers and politicians.	• To build and maintain patient, public and professional engagement and trust in health data uses. • Clear, consistent and accessible information about data uses for all stakeholders, with additional detail for those who are interested to know more.	• Inconsistent and/or unclear information from different organisations concerned with health data is a barrier to understanding and trust. • Notions of selling data for profit are unattractive to many patients, members of the public, health professionals and researchers. Undue emphasis on these damages trust in data use. • Lack of clarity on mechanisms for opt-out.	• Ongoing meaningful engagement and partnership with health professionals, (especially GPs) as well as patients and public, essential. • Consistent narrative from relevant national organisations delivered in different ways to resonate with multiple segments of society. • Focus must be on health, wellbeing and economic benefits from wide range of data uses for all patients, public and health professionals. • Trust increased if people can easily access their own health data. • Promote awareness (especially among policymakers and politicians) that: - economic gains come from increased health and wellbeing via increased productivity across all groups (by age, ethnicity, geography, deprivation etc). - pre-competitive investment from private sector in health data infrastructure (by contrast with selling data) can bring benefit for all. • Single consistent, logical, centralised, readily accessible system for opt-outs that does not impose a burden on busy GPs. • Learn from and use experience of organisations that specialise in providing clear, transparent, consistent information.[376]

376 E.g. Understanding Patient Data researches and provides balanced and accurate information on health data for patients, public, health professionals and others.

Appendix 10
Priority data requirements

Data	Characteristics	What is needed?	Why is it needed?	What are the main barriers?	How can they be overcome?
General practice data.[377]	• National. • NHS origin. • Generic. • Structured.	• Centrally coordinated, national collection of comprehensive, coded, near-real-time data from general practice systems. • In each of the 4 nations. • Linkable to other national data, as well as to population- and disease-based cohorts, clinical trials and other clinical studies.	• Replace inefficient, overlapping, costly, multiple general practice data extracts with a single system, for: • Linkage to other national data to identify people for screening, vaccination or national research. • Inclusive national and regional analyses of the causes, distribution, treatment, consequences and costs of all health conditions, whether managed in hospital or not. • Integrate linked data into research databases for efficient, cost-effective characterisation and long-term follow-up of participants.	• GP concerns about data responsibilities and potential liability. • Insufficient transparency about previous initiatives (e.g. on system design, proposed data uses, safeguards to protect privacy or prevent data misuse, opt-out mechanisms) led to discussions being dominated by concerns about risks rather than benefits. • Lack of satisfactory national solution(s) has led to (too) many general practice data access initiatives, which do not individually or collectively fulfil all requirements. • Lack of positive incentives for GPs and general practices.	• Relieve general practices of responsibility and liability for central data collection – e.g. via revisions to GP contract and/or Secretary of State directions to facilitate participation of all general practices. • Provide clear and accessible information on: proposed system, full range of beneficial uses, privacy protection, straightforward opt-out mechanism, and how decisions about data access and use (especially those that may generate profit) will be made. • Align and enhance existing national secure data access infrastructure and processes for linked general practice data. • Engage GPs in system design, and provide positive incentives to ensure their support.[378]

377 See also Chapter 3, section 3.1.2.

378 E.g., the healthcare profession and independent healthcare policy organisations have major concerns about current and worsening limitations in healthcare workforce capacity to meet increasing demands on the NHS. Better enabling access to general practice data, linked to other sources, to properly describe and understand the scale and sources of increasing GP workload would help to make the case for proposed solutions. See https://www.health.org.uk/news-and-comment/charts-and-infographics/understanding-activity-in-general-practice-what-can-the-data-tell-us and https://www.bma.org.uk/media/4316/bma-medical-staffing-report-in-england-july-2021.pdf.

Data	Characteristics	What is needed?	Why is it needed?	What are the main barriers?	How can they be overcome?
Hospital emergency dept. and admissions data.[379]	• National. • NHS origin. • Generic. • Structured.	• Enhance existing national data collections with granular, coded, near-real-time data. • In each of the 4 nations. • Linkable to other national data, as well as to population- and disease-based cohorts, clinical trials and other clinical studies.	• Weekly or daily provision of granular diagnostic and procedural data would allow rapid detection of signals for many purposes, e.g.: • Automated safety monitoring of medicines, vaccines and devices. • Integrating linked data into clinical trials and observational studies, for efficient, cost effective, participant characterisation and follow-up .	• Most hospitals do not currently implement systems for point of care clinical data coding. • National data on admitted hospital episodes is not reported until after hospital discharge (or death). • Central checking, curation and onward provision of national hospital episodes data are too slow. • Insufficient positive incentives for hospitals.	• National adoption of efficient systems for clinical coding at point of care (e.g. on the ward round). • Increase frequency of mandatory central reporting to weekly, including during hospital admissions. • Increase capacity and efficiency of central data curation and data provisioning processes. • Incentivise participation by ensuring benefits to hospitals (e.g. timely provision of data from national systems required for local intelligence).
Hospital outpatient data.[380]	• National. • NHS origin. • Generic. • Structured.	• Include diagnostic and procedural codes in national hospital outpatient episodes data. • In each of the 4 nations. • Linkable to other national data, as well as to population- and disease-based cohorts, clinical trials and other clinical studies.	• Reduce reliance on general practice data for information on diagnoses and procedures in hospital outpatients. • Linkage to other national data to identify people for screening, vaccination and national research studies. • Inclusive national and regional analyses of the causes, distribution, treatment, consequences and costs of all health conditions, whether managed in hospital or not. • Integrate linked data into research including clinical trials for efficient, cost-effective, participant characterisation and follow-up.	• Diagnostic and procedural coding of outpatient episodes is optional and only included in 3–4%. • Central checking, curation and onward provision of national hospital episodes data too slow. • Insufficient positive incentives for hospitals.	• National adoption of efficient systems for coding diagnoses and procedures in outpatient settings. • Increase efficiency and capacity of central data curation and provisioning processes. • Incentivise participation by ensuring benefits to hospitals (e.g. as above).

379 See also Chapter 3, section 3.1.4.

380 See also Chapter 3, section 3.1.4.

Data	Characteristics	What is needed?	Why is it needed?	What are the main barriers?	How can they be overcome?
Medicines data.[381]	• National. • NHS origin. • Generic. • Structured.	• National data on medicines prescribed and dispensed in hospital, and on high-cost medicines, needed to complement data on medicines prescribed and dispensed in the community. • In each of the 4 nations. • Linkable to other national data, as well as to population- and disease-based cohorts, clinical trials and other clinical studies.	• Population-wide data on medicines, irrespective of where they are prescribed and dispensed, is crucial for: • Monitoring the effectiveness and safety of medicines. • Monitoring adherence to national guidelines on medicines use. • Following national, regional and local prescribing trends. • Ascertaining many health conditions (e.g. diabetes, inflammatory arthritis) for screening and vaccination invitations, and to enhance clinical trials and other research studies on the causes, prevention and treatment of disease.	• Not all hospitals yet have an EPMA[382] system for inpatient prescribing and dispensing data. • Not all existing EPMA data have been incorporated into national data collection systems. • No nationally supported system for the collation of high-cost drugs data in England.	• Complete rollout of EPMA systems across all UK hospitals. • Reinvigorate efforts to incorporate data from all hospital EPMA systems into national data collection. • Reinvigorate efforts to collect and collate a regularly updated dataset on all high-cost drugs.

381 See also Chapter 3, section 3.1.5.

382 EPMA: electronic prescribing and medicines administration.

Data	Characteristics	What is needed?	Why is it needed?	What are the main barriers?	How can they be overcome?
Laboratory data.[383]	• National. • From NHS. • Generic. • Structured.	• Agreed UK-wide, internationally relevant terminology for laboratory tests and results. • National system for access to an evolving subset of laboratory test and result information. • In each of the 4 nations. • Linkable to other national data, as well as to population- and disease-based cohorts, clinical trials and other clinical studies.	• Population-wide laboratory data system, starting with a subset of the most commonly used tests, would enable: • Monitoring of adherence to national laboratory testing guidelines. • Setting common reference ranges. • Reduced duplicate testing. • Improved ascertainment of many health conditions (e.g. kidney and liver disease) for screening and vaccination and to enhance clinical trials and other research studies on the causes, prevention and treatment of disease.	• Laboratories currently use multiple different LIMS, middleware patches, terminologies and reference ranges. • The thousands of laboratory tests and billions of rows of data generated might make the task of generating a national system seem unmanageable.	• Reinvigorate existing efforts on agreed terminologies and reference ranges. • Start with subset of most commonly used tests to make the task more manageable and demonstrate early benefits. • Review and rationalise approach to commissioning and purchasing of laboratory computer systems, moving towards a smaller number of more interoperable systems. • Learn from and enhance successful models of national laboratory infrastructure, which include GEL genomic sequencing services and national microbiology data systems. These compare favourably to distributed, regional models for the generation of and secure access to linkable data.

383 See also Chapter 3, section 3.1.6.

Data	Characteristics	What is needed?	Why is it needed?	What are the main barriers?	How can they be overcome?
National audits and registries.[384]	• National. • NHS origin. • Domain-specific. • Structured.	• Wealth of NHS data collected in the many (estimated >100 in England) national audits and registries should be securely accessible to support a much wider range of beneficial uses than health service audit and quality improvement. • In each of the 4 nations. • Linkable to other national data, as well as to population- and disease-based cohorts, clinical trials and other clinical studies.	• These data provide detail on specific health conditions, complementing information in generic datasets. They can and should be used to: • Link to and enhance multiple research studies. • Increase data quality through its wider use. • Identify gaps and reduce duplication across parallel data collection efforts.	• Complexity: there are many national audit and registry data collections, involving multiple organisations in commissioning, funding, data controllership, data processing and data access decisions. • No comprehensive catalogue of these collections. • No transparent, national system to prioritise or coordinate secure data access and linkage for analyses bringing significant wider benefits. • No coordinated national system to prioritise linking these collections to other data, enabling investigation of data gaps and potential duplication of data collection effort.	• Create a comprehensive catalogue of all national audit and registry data collections. This should include, for each collection, a data dictionary, information on organisations involved and their roles. • NHS England should work with the many partner organisations involved to set priorities for a coordinated, streamlined route for secure data access and linkage, preferably via NHSE central access process, to support analyses that benefit patients and public.[385] • Analyses of linked data should investigate gaps and duplication of data collection to inform future data collection priorities.

384 See also Chapter 3, section 3.1.12.

385 Prioritising this work should be based on several factors, including: (i) importance of the health condition(s) addressed by the audit or registry, e.g. with respect to incidence, prevalence, mortality, morbidity, or health, care and societal cost burden; (ii) lack of adequate data from other accessible sources about the health condition(s); (iii) demand from wide range of potential users (but beware using volume of data requests as a proxy for demand, as many potential users may not know about the data or how to request it, while others may not have requested access because of a belief that the access process will take too long or be unsuccessful) (iv) whether or not the data are already held within NHS England central systems or need to be obtained from an external organisation; (v) capacity and willingness of the data controller and/or processor to provide the audit/registry data for access and linkage, e.g. by regular supply of updated data to NHS England; (vi) ease of processing and curating the data centrally (e.g. within NHS England) prior to providing as a dataset for access and linkage (noting that some audit and registry datasets are particularly well managed and curated by the audit/registry provider and so less challenging to handle within central systems).

Data	Characteristics	What is needed?	Why is it needed?	What are the main barriers?	How can they be overcome?
Screening data.[386]	• National. • NHS origi.n • Domain-specific. • Structured.	• A transparent national process for secure access to national screening programme datasets. • In each of the 4 nations. • Linkable to other national data, as well as to population- and disease-based cohorts, clinical trials and other clinical studies.	• Access to national screening data on screening programme invitees, uptake, and initial screening results, linked to health outcomes and other data, to enable: • Evaluation of screening impact on health outcomes in practice. • Evaluation of the potential impact of more targeted screening programme inclusion criteria.	• Data for the 11 national screening programmes is held and processed by a range of different public and commercial organisations. • Data access requests for English screening programme data are not dealt with by NHSE's central data access request process[387] but by a separate process overseen by six NHSE screening Research Advisory Committees, with no clear guidance on obtaining linkage to other health data.[388]	• Screening datasets should be included within NHSE's centralised data access process to facilitate access, linkage and generation of insights. • Increased efficiency and capacity of central data curation and provisioning will be needed. These could be facilitated through partnerships with motivated external experts. • Screening experts should represent national screening programmes in data access decisions.
Social care data.[389]	• National. • Originates beyond NHS. • Generic. • Structured.	• National social care datasets, including non-publicly funded care, covering all age groups, and able to identify and track care home residents. • Person-level and frequent (monthly or weekly for close to real-time insights). • In each of the 4 nations. • Linkable to other national data, as well as to population-and disease-based cohorts and clinical trials.	• Access to and analyses of social care data linked to health data could generate insights on delays in care pathways and their costs, inform service planning, assess inequalities in the provision of care, and look at the impact of different types of care on health outcomes. This is currently not happening.	• Incomplete digitisation across social care limits potential for frequent, person-level, national data collection. • Lack of agreed UK-wide national core social care data items. • Linkage potential limited because NHS number (or CHI in Scotland) not always included. • No coverage of non-publicly funded care. • Social care data from local authorities is provided to NHSE under auspices of DHSC for adults and to the Dept of Education for children.	• Local government must receive ongoing funding and guidance to achieve digital maturity rapidly across social care sector. • Health and social care and education depts across the four nations to agree on core data items. • Health and social care departments to mandate inclusion of NHS number or CHI in adult social care data. • National and local government to mandate inclusion of data on non-publicly funded care in national collections (using legal mechanisms if required).

386 See Chapter 3, section 3.1.8.

387 See https://digital.nhs.uk/services/data-access-request-service-dars.

388 See https://www.gov.uk/guidance/nhs-population-screening-data-requests-and-research.

389 See Chapter 3, section 3.2.2.

Data	Characteristics	What is needed?	Why is it needed?	What are the main barriers?	How can they be overcome?
Other cross-sectoral data.[390]	• National. • Originates beyond NHS. • Generic. • Structured.	• Secure, streamlined access to health-relevant administrative data from non-NHS and social care sources, linked to data from the health and care system.	• To enable policy relevant insights on the wider determinants of health and wellbeing, and the consequences of mental and physical ill health, including a deeper understanding of the causes and consequences of health inequalities.	• Lack of clarity on legal gateways for sharing and linkage of health and care to other sources of administrative data. • Lack of streamlined mechanisms for sharing and access of data between NHSE and ONS. • Lack of agreed criteria for accreditation of secure environments holding health and care data.	• Review inclusion of health and care data in Digital Economy Act, including consultation with healthcare profession. • NHSE/ONS partnership to develop streamlined mechanism for interorganisational data sharing and access. • UK-wide system for standards and accreditation of secure environments holding data from the health and care system.
Imaging data.[391]	• Regional/national. • NHS origin. • Generic or domain-specific. • Unstructured.	• Large-scale population-based imaging resources based on routine NHS imaging activity. • In each of the 4 nations. • Linkable to other national data, as well as to population- and disease-based cohorts.	• Such resources, securely accessible and linked to other health data, needed for: • Discussion and distributed reporting for better clinical care. • Developing and testing automated imaging processing and analysis tools (many AI-based), prior to evaluation and implementation in live NHS systems. • Research studies to understand the impact of imaging and interventional radiology procedures on subsequent health outcomes and how structure and function of body organs influence subsequent health.	• Main challenges are technical: • Complex, unstructured nature and relatively higher volume of imaging data (cf. most structured, coded data) require special approaches for storage, transfer, format standardisation, de-identification, security, analysis, and linkage to other data sources. • Poor interoperability across the many and varied computer systems for handling NHS imaging data.	• Learn from successful examples, e.g. Scottish Medical Imaging resource for radiology images and National Pathology Imaging Co-operative for histopathology images, that have started to address key technical challenges. • Use such examples to inform further efforts to enable at scale access to and linkage of large collections of well-curated NHS images, with realistic but ambitious goals for expanding this capability across the UK.[392]

390 See Chapter 3, section 3.2.3.

391 See Chapter 3, section 3.1.7.

392 E.g., now that secure access to and linkage to other national health data of all imaging covering 10 years of NHS radiology activity across Scotland (population 5.5 million) have been demonstrated, an advance would be to demonstrate similar capability for larger populations (e.g. >10 million), together with capability to support analyses (e.g. developing and testing new AI analysis solutions) run across secure data environments in different parts of the UK holding routine NHS imaging data.

Data	Characteristics	What is needed?	Why is it needed?	What are the main barriers?	How can they be overcome?
Other granular unstructured data.	• Regional.	• Widespread use of tools to securely interrogate very large volumes of unstructured data in health and care system (especially free text), and to capture structured data outputs from these.	• 80% of healthcare data is in unstructured form, containing a wealth of largely untapped, granular information that could enhance understanding of health wellbeing and disease. • Automated coding in real time of electronic patient records would transform quality, depth and timeliness of structured information e.g. in national hospital episode statistics.	• Until recently, health and care systems were insufficiently digitally mature to enable implementation of the relevant technologies. • Limited funding/resources • Insufficient data, tech and informatics expertise to commission wisely.	• Rapid rollout of EPRs in recent years means that widespread adoption of relevant tools across many primary and secondary healthcare organisations should now be possible. • Support for appropriate commissioning of cost-effective solutions that will create good return on investment. • Potential for coordinated partnerships between multiple trusts in commissioning to achieve economies of scale.

Appendix 11
Options for a national system for general practice data in England

Option	Pros	Cons	Additional comments
1. Comprehensive, population-wide, structured, coded data, extracted from general practice computer systems into NHSE systems.	• Technical solution for extract and regular near-real-time updates already developed so could be implemented rapidly. • Cost effective and efficient – could replace multiple GPES[393] extracts. • Would create a national general practice data resource capable of fulfilling all beneficial use cases (Table 7.1).	• Similar programmes in 2014 (care.data[394]) and 2021 (GPDPR[395]) failed or stalled due to concerns from GPs, privacy groups, patients and public about – burden (workload and potential liability) on GPs. – privacy and security. – potential data misuse. – complexity of opt-out.	• Many concerns have been addressed since GPDPR was paused, but still need: – ongoing public, patient and professional engagement. – clear and simple opt-out system. – transfer of data responsibility and liability from general practices to NHSE (e.g. via SoS direction.) – possible revisions to GP contract • Could incorporate OpenSAFELY interface within NHSE SDE for some uses (see option 4 in this table).
2. Expand CPRD.[396]	• Long standing, widely used data collection and provisioning system with experienced team. • Existing large user base across academia and industry. • Hosted and used extensively by MHRA for medicines safety monitoring. • Provides trial recruitment and follow-up services via involved general practices. • Cost-recovery pricing model means it does not rely on external funding.	• Relies on general practice opt-in, with incomplete population coverage (around 30%) despite attempting to reach full population coverage since 2011. • Relies on transfer of data from NHSE and other sources to enable linkages. • Linkages with other data (via NHSE/others) cumbersome and time-consuming. • Lags behind real time. • Considered expensive, especially among university-based researchers. • As currently set up, cannot support all use cases required of a national general practice data system (see Table 7.1).	• Could incorporate CPRD services within proposed national health data service (section 7.1.2) to broaden population and use case coverage. Would need to ensure ongoing ready accessibility for MHRA to maintain and enhance existing capability for medicines and devices safety monitoring. Would remove the need for CPRD to maintain its own separate SDE, reducing costs for service users, while retaining substantial CPRD expertise. • Since it was established in 2011, the coverage of CPRD has expanded from 10% to around 30% of English general practices. It seems vanishingly unlikely that general practice opt-in mechanisms will expand coverage to anywhere near 100% without SoS[397] direction +/- changes to GP contract.

393 General Practice Extraction Service: https://digital.nhs.uk/services/general-practice-extraction-service.

394 See https://www.england.nhs.uk/wp-content/uploads/2015/07/care-data-Quick-reference-guide-v1.0.pdf.

395 General Practice Data for Pandemic Planning and Research: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/general-practice-data-for-planning-and-research.

396 Clinical Practice Research Datalink: https://www.cprd.com/join-growing-network-practices-contributing-cprd.

397 SoS: Secretary of State (for Health).

Option	Pros	Cons	Additional comments
3. Expand RCGP Research and Surveillance Centre (RSC).[398]	• Long-standing general practice-based disease surveillance system (established 1957). • Provides access to de-identified data via a SDE. • Provides trial recruitment and follow-up services via involved general practices. • Provides disease surveillance services (including via bio-samples) for UKHSA.[399] • Strongly supported by RCGP.	• Relies on general practice opt-in with incomplete population coverage (around 25-30%). • Relies on transfer of data from NHSE and other sources to enable linkages. • Linkages with other data (via NHSE/others) cumbersome, time-consuming and lag behind real time. • As currently set up, cannot support all use cases required of a national general practice data system (see Table 7.1).	• Could incorporate most/all RSC services within proposed national health data service (section 7.1.2) to broaden population and use case coverage. Would remove the need for the RSC to maintain its own separate SDE, reducing costs for public funders and service users, while retaining substantial RSC expertise. • Since it was established in 1957, the coverage of RSC has expanded to around 25-30% of English general practices. It seems vanishingly unlikely that general practice opt-in mechanisms will expand coverage to anywhere near 100% without SoS direction +/- changes to GP contract.
4. Implement OpenSAFELY capablities within NHSE SDE.	• Additional privacy/security layer meets with approval of RCGP, BMA and privacy groups • Reproducible data curation and analysis pipelines would add to existing capabilities in NHSE SDE, enhancing efficiency	• OpenSAFELY interface not suitable for all analyses that need access to comprehensive, de-identified record level data (e.g. some AI methods). • Would still need secure transfer of data out of NHSE for certain purposes, in particular 'consented' research cohorts and trials.	• Could be provided in combination with option 1 in this table.
5. Explore data within general practice computer systems via OpenSAFELY and extract to NHSE (or other secure settings) as needed.	• OpenSAFELY implemented within general practice systems meets with approval of RCGP, BMA and privacy groups. • Avoids extraction of comprehensive coded general practice data from commercial general practice computer systems.	• Significant scalability and efficiency challenges for multiple queries and extracts by NHSE analysts and researchers from academia and industry.[400] • Implementation of some analyses (e.g. AI methods) would need access to comprehensive, de-identified record level data. • Dependent on ongoing agreements for hosting of OpenSAFELY interface and access for multiple users within all commercial general practice system suppliers' systems for English general practices (currently two suppliers cover almost all practices but there may be more in the future). • Relies on ongoing flow of data from NHSE (and other data custodians) into each of the commercial general practice system suppliers' systems.	

398 RCGP Research and Surveillance Centre: https://www.rcgp.org.uk/clinical-and-research/our-programmes/research-and-surveillance-centre.

399 UKHSA: UK Health Security Agency.

400 E.g. 100s to 1000s of research projects for UK Biobank alone.

Cathie Sudlow was supported in carrying out her review by a team at Health Data Research UK.

Cite this report as: Sudlow, CLM (2024). Uniting the UK's Health Data: A Huge Opportunity for Society.

https://doi.org/10.5281/zenodo.13353747