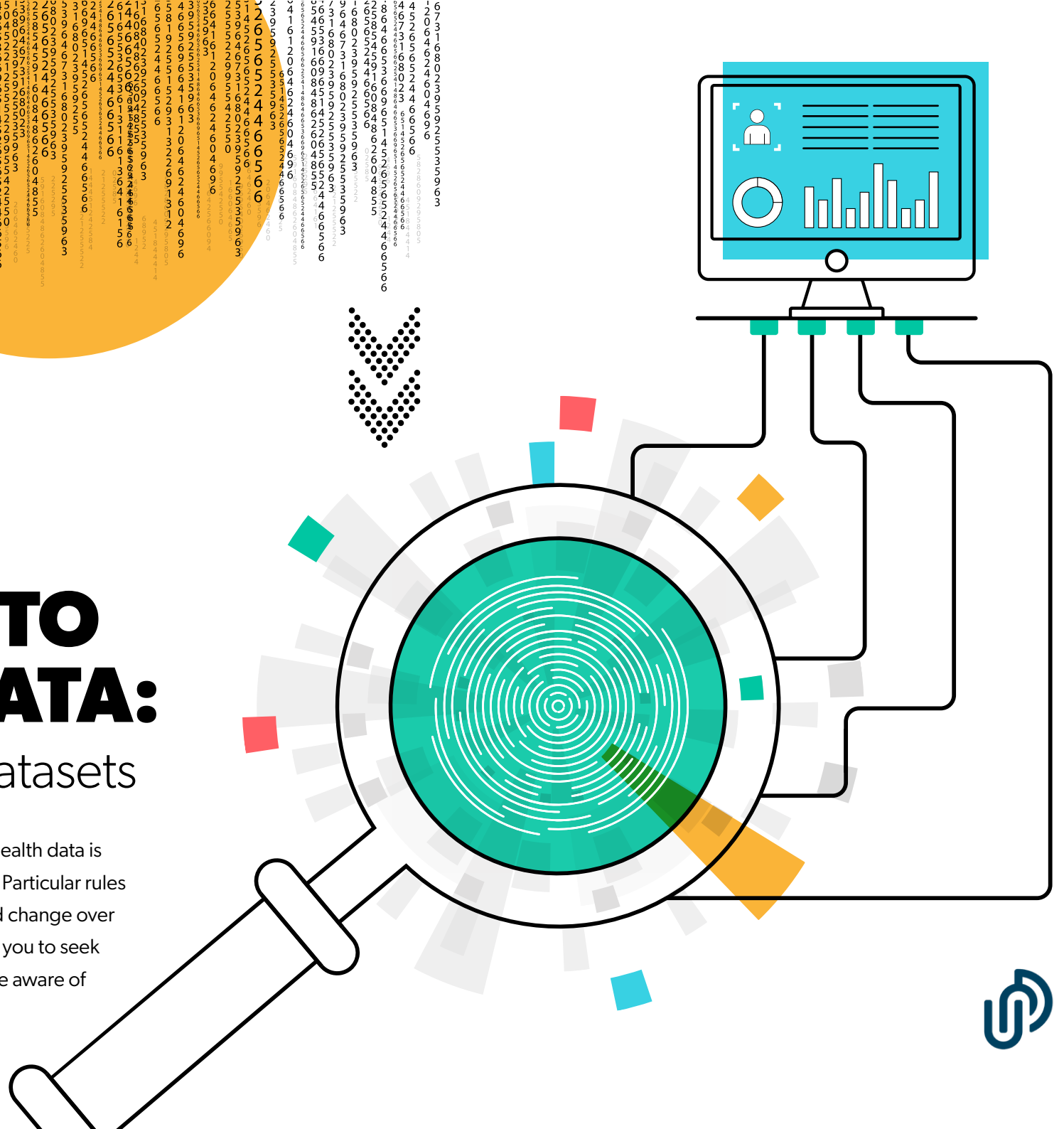




WHAT HAPPENS TO HEALTH DATA:

A guide to large datasets

This guide is an introduction to the way that health data is gathered, stored and used in the NHS today. Particular rules and guidelines may vary in different areas and change over time. However, we hope this guide will allow you to seek out further information and help you feel more aware of the data that is kept about you.



I'D LIKE TO KNOW...

...what this guide is about	03
...what health data is, and why it is collected	05
...about the potential benefits of large health datasets	08
...about the potential harms of large health datasets	11
...how my privacy is protected	14
...who can and cannot access health data	19
...how health data is stored in a secure way	21
...what is the future for large health datasets	24
...where I can find more information about health data	25

WHAT IS THIS GUIDE ABOUT?

The healthcare you receive from the National Health Service (NHS) relies on the information that is recorded and kept about you. That information is often held in large collections of digitally stored health data. These large collections (or 'sets') of health data are gathered from everyday interactions between people and healthcare services. Large datasets have the potential to improve many aspects of healthcare, and are the main focus of this guide.

All health services need data to work effectively, and always have done. Every decision or recommendation for your health is based, in part, on your medical history. This history includes previous discussions with healthcare workers and test results. It may also include other information such as allergies, lifestyle habits and family medical history.

However, data is not only used to make healthcare decisions for individuals. Data is also essential in researching new treatments and planning healthcare for whole populations.

Healthcare research and planning often use data gathered from limited numbers of patients. This data can be used alongside expert knowledge and experience to draw conclusions and make decisions. Larger datasets – that include data gathered from many more patients – have the potential to improve the evidence we have available.

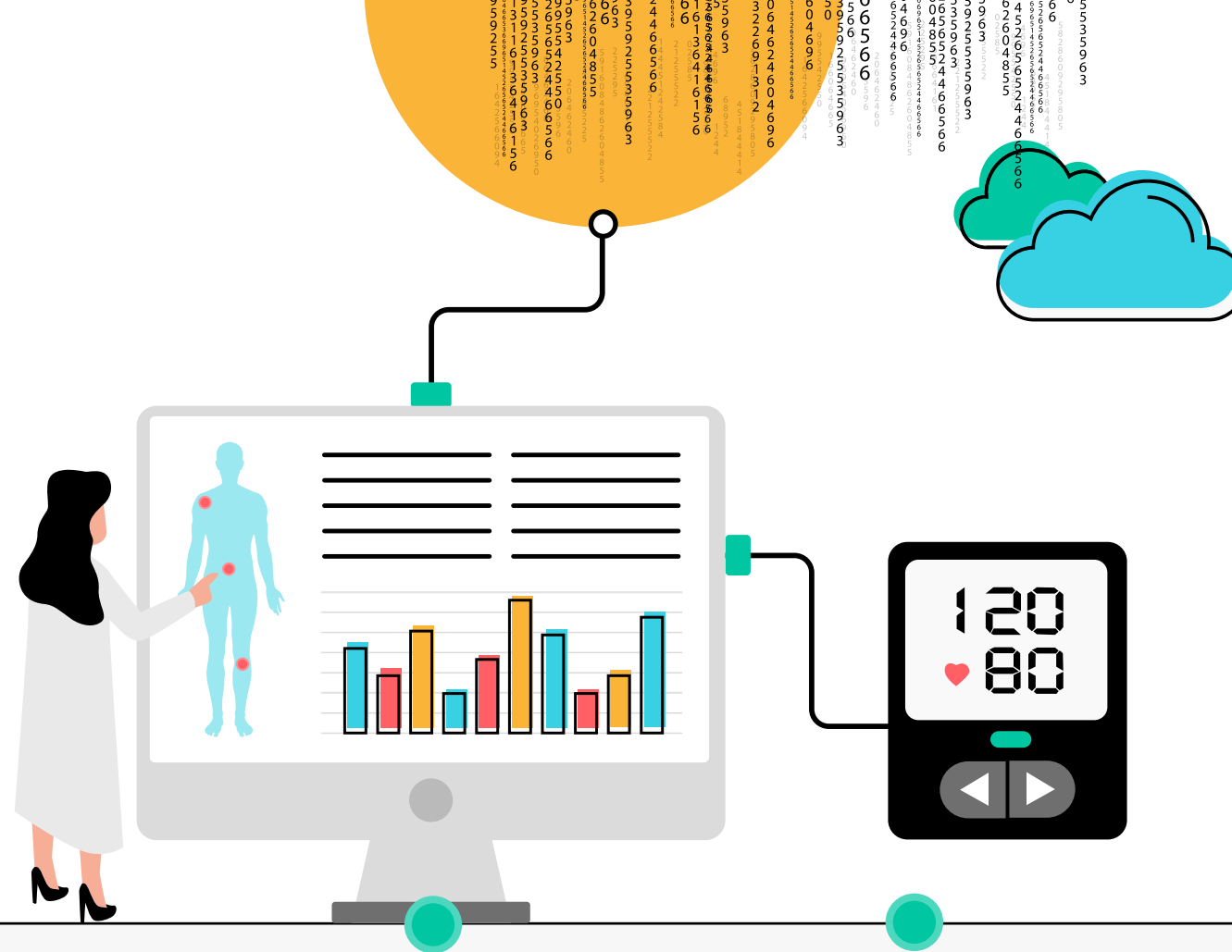
Storing information digitally: promises and pitfalls

Digital data is information that is stored electronically, rather than on paper. Digital technology means that more health data is now stored than ever before. This brings great opportunities, especially when data is used for research and planning at a national level.

However, the digital storage of health data – and its use for research and planning – has also brought legitimate concerns about how that data is handled, including worries about personal privacy.

COLLECTING DATA

This example is hypothetical and illustrative: it shows only one way in which health data might be collected.



You visit your general practitioner's (GP) surgery for a general health check-up, including having your blood pressure taken.

The doctor asks questions about your medical history, your family history and whether you are currently experiencing any symptoms. They make notes about this discussion on their computer.

Your GP then physically examines you and takes your blood pressure. The findings are again added to the computer notes.

After a brief discussion, you and your GP decide it would be a good idea for you to have some blood tests, which will be done at the nearest hospital. You agree to make another appointment with your GP once these results are back.

WHAT IS HEALTH DATA?

Health data is any information about a person's physical or mental health: in the past, at present or in the future.

This means that health data may come from any interaction with the healthcare system, for example during an appointment with a GP, or with a nurse or doctor in a hospital. Health data may also be gathered from medical devices and from diagnostic tests (for example, blood tests or genetic tests).

This enormous range of information is recorded and stored in many ways, using many different types of computer system. Some health data will be fairly straightforward numbers (for example, your height or weight). Other information will be stored as free-text notes (for example, comments written by your GP during a consultation).

Building large datasets, where the information from many patients is recorded in a consistent way so that it can be analysed, is a big challenge.

Before considering the possible benefits and risks of overcoming this challenge, it is worth thinking about the reasons for collecting health-related information in the first place.

1. Ibrahim H, et al. *Lancet Digital Health*. 2021;3(4):e260–e265. 2. Indirect Government Services: Choosing an IT device. Available at <https://www.nidirect.gov.uk/articles/choosing-it-device> (accessed May 2022).


2,314,000,000,000,000,000,000 bytes¹

Approximately equivalent to the storage capacity of up to

4.6 billion
typical laptop computers²



= The estimated total volume of health data in 2020

 = 1 billion

WHY IS HEALTH DATA COLLECTED?

First and foremost, health data is collected to help provide healthcare to the person from whom the data was collected.

For example, as shown in the case study, your GP will often ask questions about the reason for your appointment and about your medical history. They may perform a physical examination, and they may request further tests, such as blood tests or scans.

The results of all these investigations will be added to your notes, which are kept on an electronic health record (EHR).

This information includes your personal details – such as your name, address and NHS number – and will help healthcare professionals when they are making diagnoses and decisions relating to your care.



The storage of this data enables clinical care to be provided effectively, as all professionals involved in your care should have access to the information they need about your health.

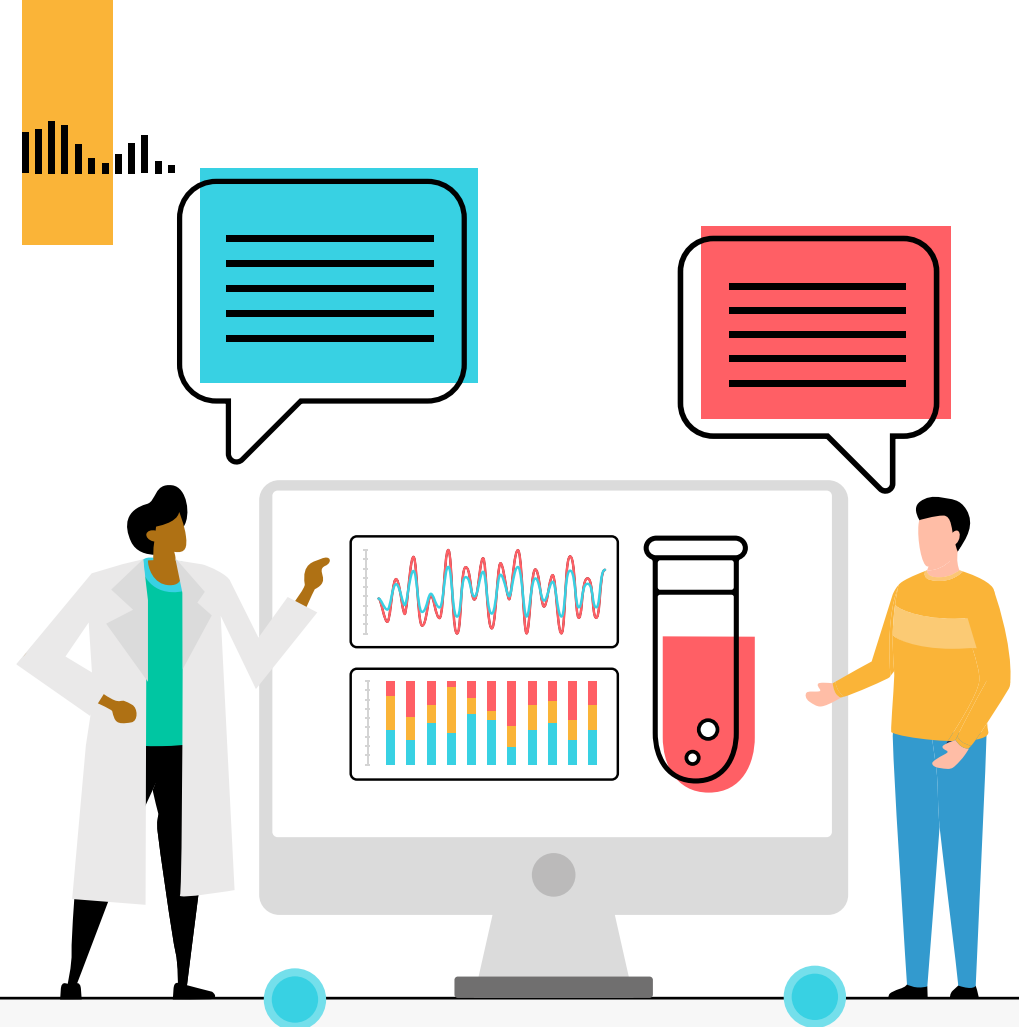
For example, the doctors at your local GP practice and in hospital departments should be able to see your history of treatments and medical problems. Tests performed in different places can be added to your EHR so that all the information is kept together.

Having your personal details in this information (including your individual NHS number) helps ensure that results are not mixed up between patients. It also allows you to be contacted about results and appointments.

These uses are known as the primary – or direct – use of health data.

THE PRIMARY USE OF DATA

This example is hypothetical and illustrative: it shows only one way in which health data might be used.



At your next appointment at your GP's surgery, you see a different doctor.

You discuss the findings from your last appointment and the results of your blood tests. You also discuss how you have been feeling since your last appointment.

Together, you agree that you do not need any new medications at the moment. You agree to make some lifestyle changes that may help to keep your blood pressure within a healthy range.

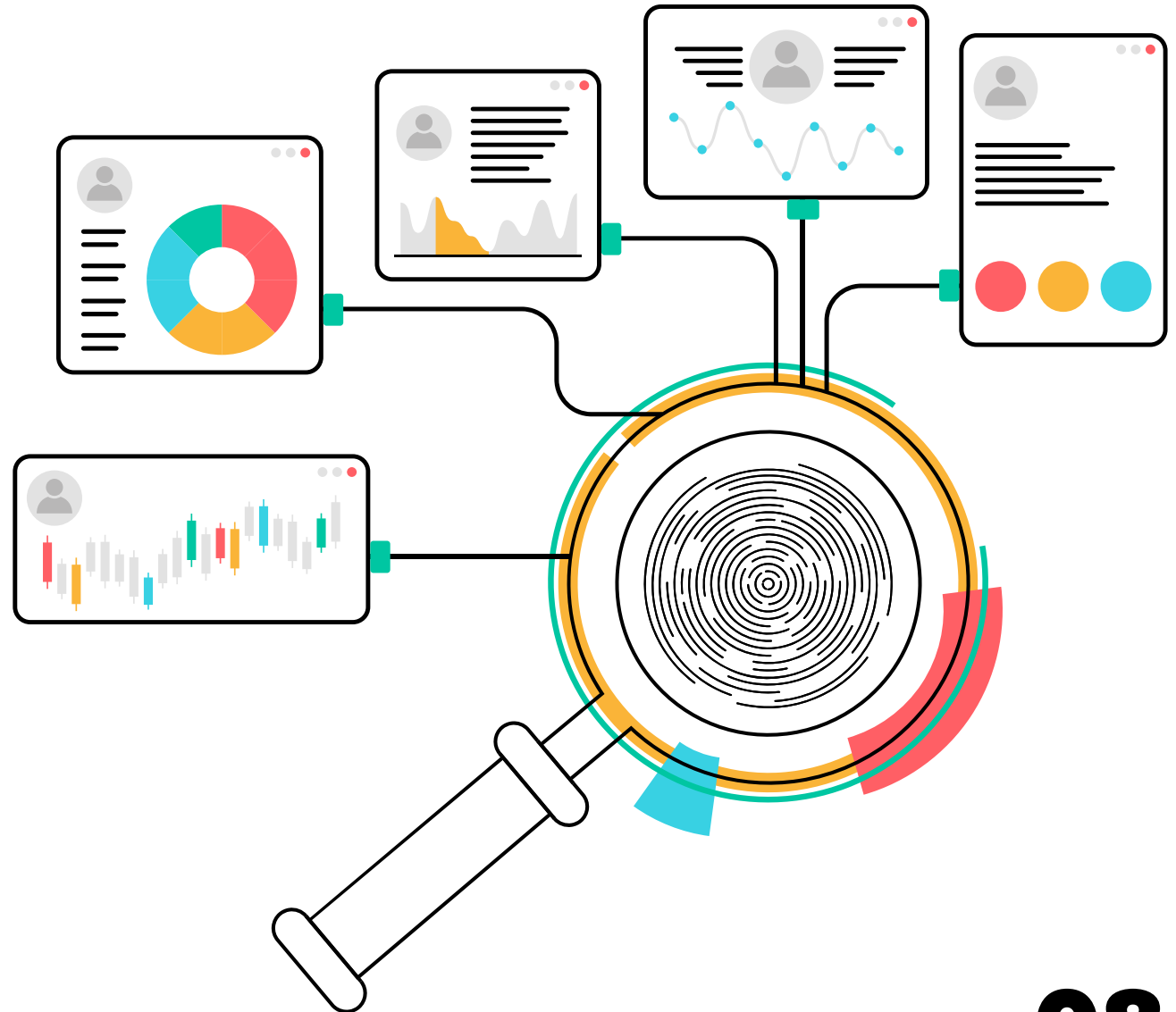
You arrange to make a follow-up appointment in a few months' time.

WHAT ARE LARGE DATASETS?

Large datasets are created when the health data of many people is gathered together. When it is stored digitally and organised properly, huge amounts of data gathered from large numbers of patients can be analysed.

This analysis can be used to improve the understanding of diseases and disability, and to develop new treatments and technologies. It can also be used to plan healthcare services for all our future needs.

When health data is used in this way – for research and planning – it is known as secondary or indirect use of data.



WHY ARE LARGE DATASETS CREATED?

The potential benefits of large datasets include:



Identifying better ways to predict and diagnose illness

When medical staff have good evidence – which can be provided by analysing lots of data – they can use it to recommend the most effective care for patients.

- Having data about large numbers of patients means that better predictions can be made. For example, predictions about health risks for specific groups or even a whole population.
- Large datasets may also allow rarer diseases to be researched. Diseases that are very unlikely to occur in a small group of patients may be seen in data from large populations.



Developing new treatments, and monitoring the safety of existing treatments

Analysing large datasets can help researchers see both the short-term and the long-term effects of treatments. Some of this work will be done by academic researchers or scientists within the NHS, but some will be done by commercial organisations such as pharmaceutical companies.



Planning services

When the NHS knows how many people have a particular health condition, and what their care needs are over time, data can be used to plan which services they will need in the future.

Data on healthcare can also reveal how resources might be better used or to plan for extreme events such as pandemics. This information can improve responses and help reduce costs across the health service.



Addressing health inequalities

When health institutions and governments have data that covers the whole population, they can detect areas or groups of people that have worse health outcomes and target resources towards those who need them most.

THE CREATION OF A LARGE DATASET

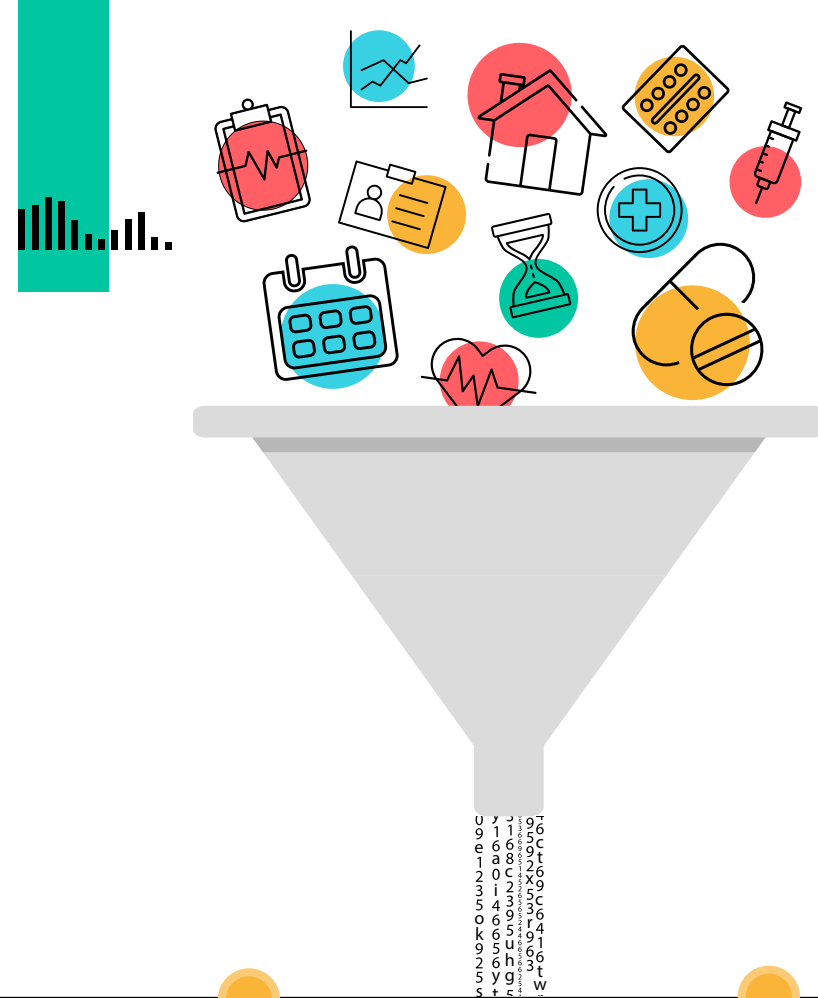
This example is hypothetical and illustrative: it shows only one way in which health data might be used.

The information from your previous GP appointments is included – along with data from thousands of other patients in the UK – in a dataset about high blood pressure (known medically as ‘hypertension’).

This dataset includes your age, your other health conditions, any medications you have been on, and the number of hospital appointments you have had in the past 5 years.

Your name, address and NHS number have all been removed from the data, and replaced with a random set of letters and numbers. This process is called ‘pseudonymisation’ and is used to reduce the risk of the researcher identifying you from the data.

You can find out how and why this process is used in the next section on [Data Privacy](#).



WHAT ARE THE POTENTIAL HARMS OF LARGE HEALTH DATASETS?

Any use of health data is entirely dependent on the quality of the data itself. To be useful, the data must have been collected, managed and organised properly.

To help answer questions about the overall population, the data must also have been collected from a large and representative group of people.

This is important in the example of health inequality. If there is inaccurate or incomplete information about people's needs, this could result in too little care being provided for groups who are already disadvantaged.



WHAT ARE THE POTENTIAL HARMS OF LARGE HEALTH DATASETS?

The potential harms of gathering health data can include:¹



Privacy breaches

The potential for unjustified or unauthorised intrusions into a patient's personal or private information.



Discrimination and stigma

Data could be used to characterise individuals or groups in ways that create – or add to – disadvantages. This can happen intentionally or unintentionally.



Disenfranchisement and disempowerment

The concern that patients do not have sufficient control over their data. It may be a significant problem if there is a lack of engagement with the public so that people are not aware of how data about their health might be used.



Exploitation

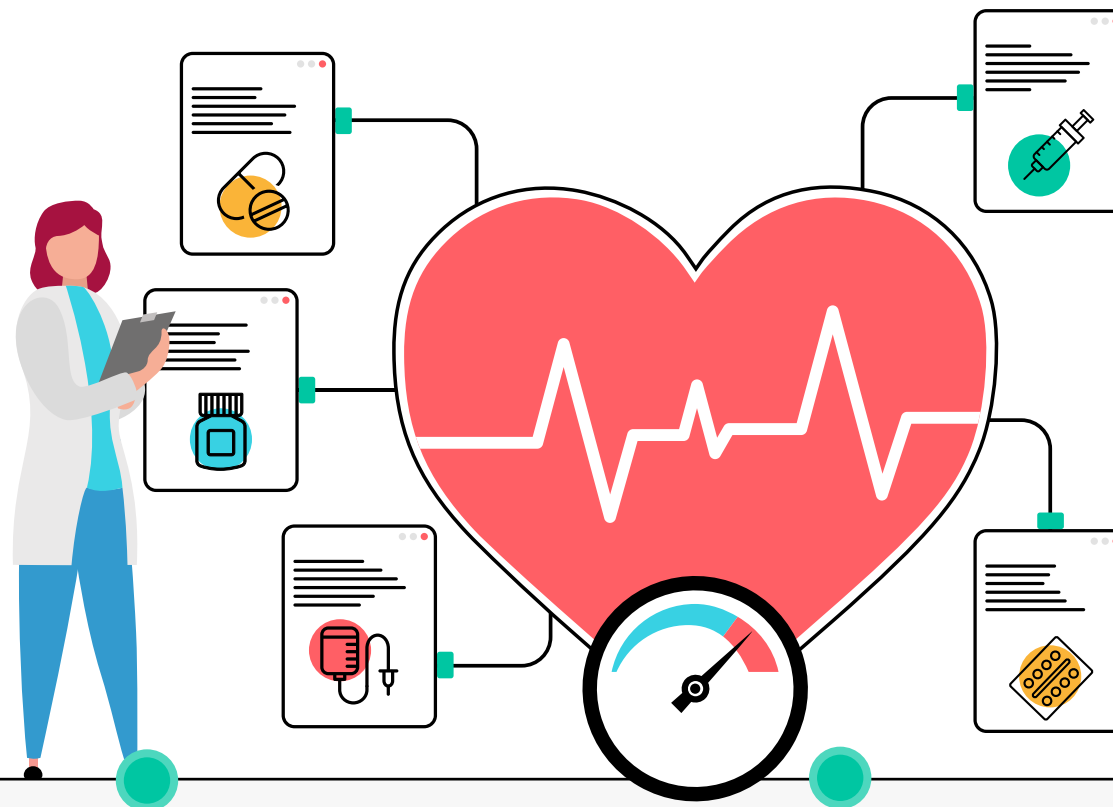
There have been concerns that the NHS does not benefit sufficiently from data that is shared with commercial organisations (such as pharmaceutical or data technology companies).

However, it has also been argued that the benefit is achieved through improvements in clinical care. Some have suggested that the NHS – or even patients themselves – should receive greater financial compensation for the use of their health data.

1. Ballantyne A. *Journal of Medical Ethics*. 2020;46(5):289–294.

USING LARGE DATASETS FOR RESEARCH

This example is hypothetical and illustrative: it shows only one way in which health data might be used.



The dataset – which includes pseudonymised information about your medical history – is being used by a researcher within the NHS. This researcher wants to find out how hypertension is monitored and managed within the health service.

They want to know what medications are being used by patients with hypertension, and how effective they are. They also want to find out if there are any links between high blood pressure and factors such as age, sex or ethnicity.

The evidence that they gain from this dataset could help improve the treatment of patients with hypertension in the future.

HOW IS MY PRIVACY PROTECTED?

The 'Five Safes' of data handling

These are a series of principles for the use of data originated by the UK's [Office for National Statistics \(ONS\)](#).

These principles were originally created for the use of any type of data. However, in recent years they have been adopted by health data organisations. These organisations include [Health Data Research-UK \(HDR-UK\)](#), the [National Institute for Health Research Design Service](#) and other secure data environments in the UK (see the [Further Information](#) section to find out more about these organisations).

These 'Safes' can be thought of as questions that should be asked by anyone who manages health data within any organisation, including the NHS. Organisations or individuals that have this role are often known as 'data controllers'.

If satisfactory answers can be given to all of these questions, then the use of the data can be considered to be secure (see the [Further Information](#) section to find out more).

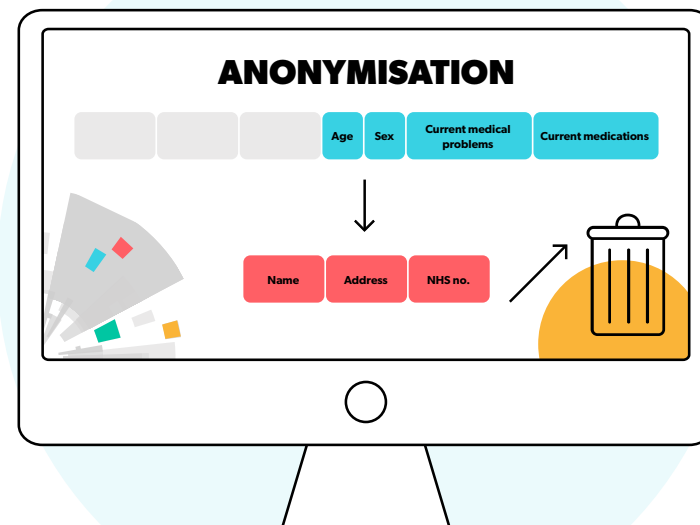
- 01 Safe data**
Does the data contain confidential information? In other words, has the data been sufficiently treated to protect your privacy?
- 02 Safe projects**
Is the use of the data appropriate, lawful, ethical and sensible? Crucially, this also includes the question: is the data being used for the public good?
- 03 Safe people**
Can the users of this data be trusted to use it in an appropriate way? Have the researchers been trained and authorised to use data safely?
- 04 Safe settings**
Is the facility or environment in which the data is held secure? Are there enough protections to prevent unauthorised use of the data?
- 05 Safe outputs**
What will be the outputs – or end results – of the data analysis, and have these been approved to ensure they do not disclose personal patient information?

HOW IS MY PRIVACY PROTECTED?

How de-identifying data can help preserve your privacy

De-identifying health data involves removing – or disguising – your personal information so that it is difficult to single you out of the dataset.

There are many ways of doing this. The method – or methods – chosen will dictate how well your private information is protected.



Anonymisation

Data is considered to be ‘anonymised’ if all identifying information (such as your name, address or NHS number) is completely removed. This means that your identity cannot be re-linked to the data in a straightforward way.

Data that is treated in this way is no longer considered confidential and, generally speaking, does not fall within data protection laws.

It is important to note that even when health data is anonymised, it may still be possible to find ways of identifying your individual, personal information. However, it would likely require special circumstances or effort. This effort may involve using other sources of information to narrow down the number of individuals that the data may be referring to.

As a result, there is some debate about when data should be considered completely ‘anonymous’.

HOW IS MY PRIVACY PROTECTED?

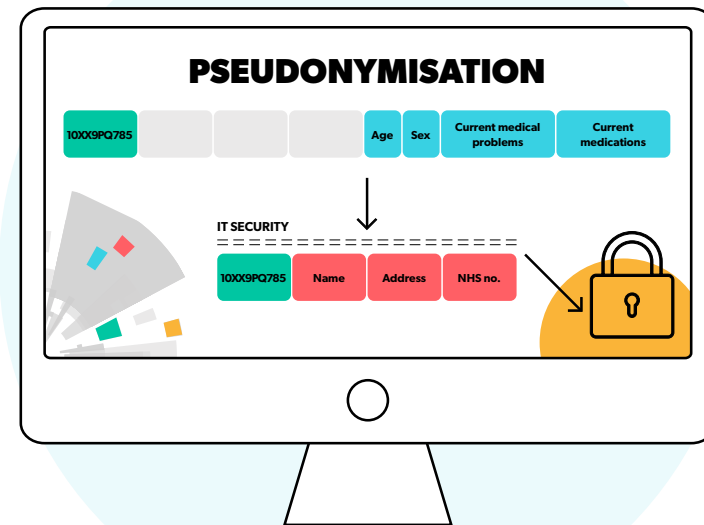


Pseudonymisation

This very common method uses a process to ‘pseudonymise’ personal information.

This means that a unique marker is used in place of identifying information (name, address, etc.). This marker is sometimes created by scrambling the identifying information, to produce a random-looking string of letters and numbers.

This unique marker does not itself reveal an individual’s identity, but distinguishes between individuals in a set of data. The original identifying information is kept separate from the dataset through technological and legal barriers.



HOW IS MY PRIVACY PROTECTED?



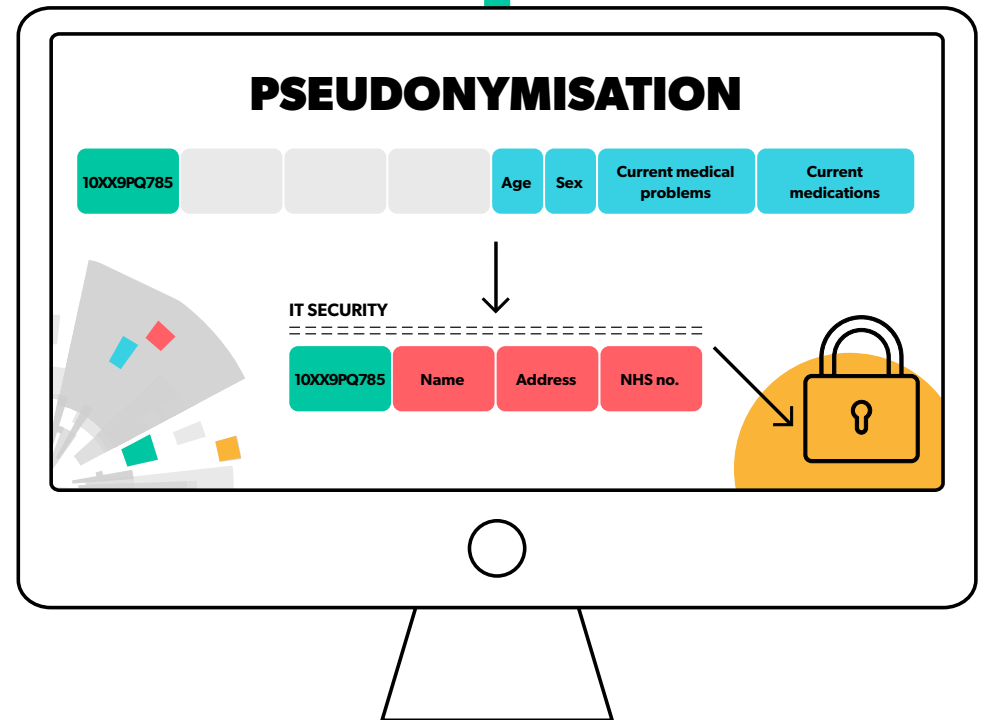
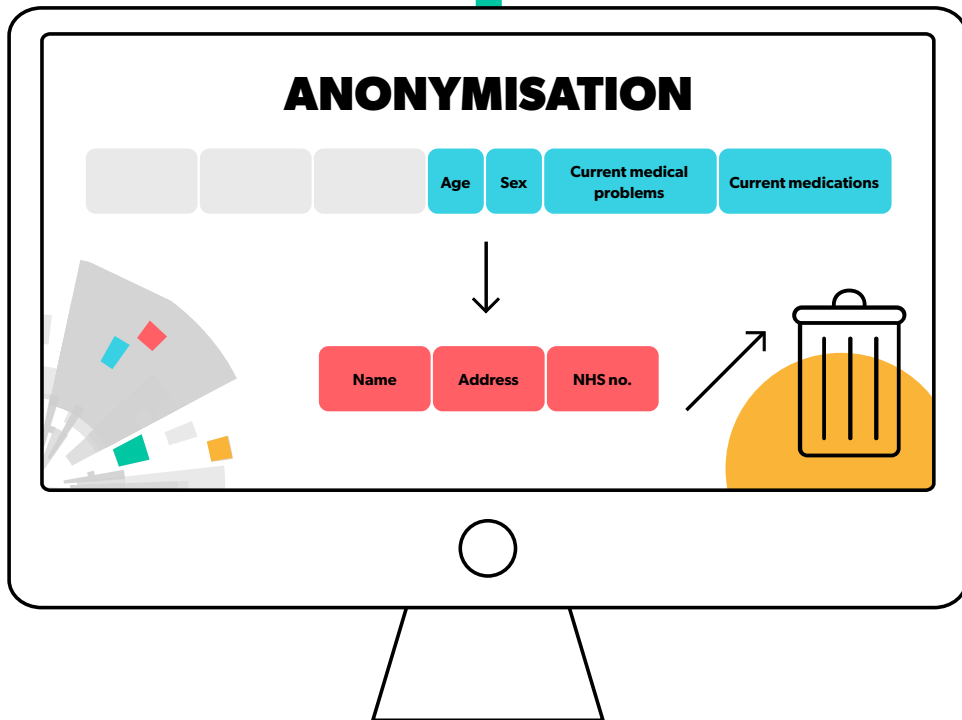
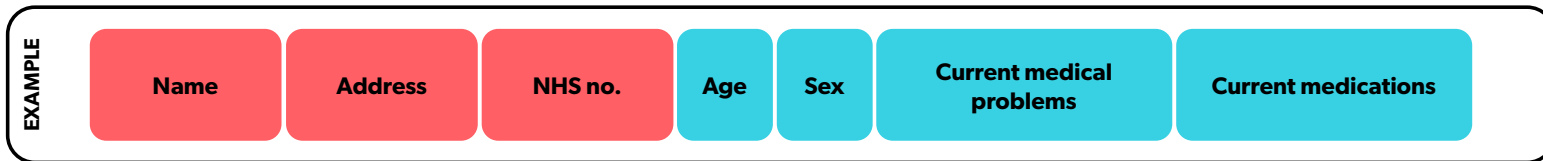
Identifying information
(e.g. name, address, NHS number)



Medically relevant information
(e.g. conditions, treatments, procedures. May also include, for example, age and sex)



CONFIDENTIAL PATIENT INFORMATION



OTHER WAYS TO DE-IDENTIFY DATA

Specific information may be replaced with something more general, for example your age rather than your date of birth, or the name of your region rather than your complete postcode. This technique is known as **'derivation'**.

Another technique involves combining the information about you and other patients into broader categories. An example of this is using age categories – such as 0–17, 18–35, 35–50, etc. – rather than individual ages. This technique is known as **'aggregation'**.

Which approach is best?

This is a very difficult question to answer, partly because there are many different ways that data can be pseudonymised or anonymised, or combined with other techniques that reduce the chance of you being identified.

These de-identification techniques can be thought of as a spectrum. At one end, there is confidential information that contains your personal details – for example, your medical notes written by your GP that contain your name and full address. At the other end of the spectrum is completely anonymous information. An example of this is the data that can be found on the [NHS Digital Dashboards](#), which has been de-identified and aggregated.

There are costs and benefits to any approach

The more anonymous data is – i.e. the less possible it is for it to be linked to your personal information – the less valuable it might be for analysis and research. For example, it may be important to know the age of individuals who most commonly have a particular condition.

WHO CAN ACCESS HEALTH DATA?



Non-profit research organisations, including universities and charities.



Commercial organisations, for example pharmaceutical companies when they are researching and developing new drugs or treatment, and technology companies that provide software or data analysis services.



Branches of national government, such as the [Department of Health and Social Care](#), and [NHS England](#).



Local authorities



Primary care networks, which are local networks that include GPs and community, social care and mental health services, as well as pharmacies and voluntary services.



Clinical commissioning groups, which are groups of GP surgeries, and **integrated care organisations**, which are groups of health and care providers in an area.

WHO CANNOT ACCESS HEALTH DATA?

For an individual or organisation to have access to NHS health data, they must demonstrate that their intended use of the data is appropriate. They also must demonstrate that the data is being used for the clinical benefit of patients, and that they will handle the data with all the necessary safeguards.

Anyone who fails to meet these criteria will not be given access to NHS health data. When making such decisions, the NHS does not distinguish between general types of organisations or individuals. For example, the decision is not affected by whether the person requesting access works for a commercial company or for the NHS itself. The only considerations are how the data will be handled and the purpose of its use.



HOW IS HEALTH DATA STORED IN A SECURE WAY?

Large datasets are stored and accessed in many different ways, and the methods will vary between different GP practices, hospitals, regional authorities and other healthcare organisations.

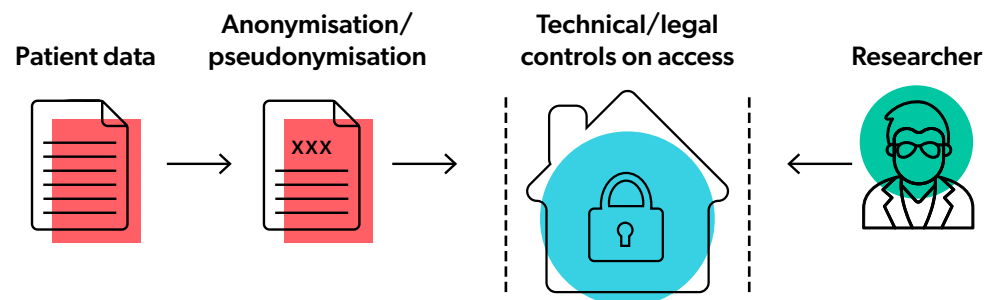
However, the UK government has stated that one particular model will become the standard across [NHS England](#). This model requires that the data is never shared outside a particular digital environment, which is known as a **secure data environment (SDE)**.

Data held in an SDE cannot be moved elsewhere, either within the NHS or beyond. Only those people who have been granted access to the data are able to log in to that digital environment.

As no data that can be linked to an individual leaves the server, and all access to the data and analysis is monitored, this approach greatly reduces the risk of data breaches or other misuse.

Currently, most use of NHS data is not within an SDE. However, their advocates argue that SDEs offer better oversight and protection against untrustworthy behaviour by individuals or organisations.

SECURE DATA ENVIRONMENT



WHAT PROTECTIONS ARE IN PLACE FOR PATIENT PRIVACY?



Technical protections

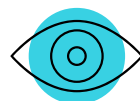
These include making changes to the data itself, so that it is much harder to identify individuals from the data (see section '[De-identifying data](#)') and creating secure data environments where no data leaves the NHS (see information above on [secure data environments](#)). Different systems will place emphasis on different aspects of these technical protections.



Legal protections

These are obligations and responsibilities placed upon any organisation or individual with access to health data. Such protections are set out in legally enforceable contracts. For example, if the [UK Information Commissioner](#)

identifies that an organisation has not complied with data protection legislation, they can impose fines of up to £17 million or 4% of global turnover (for the most serious data breaches). However, to date, such fines in the health sector are very rare.



Patient oversight

This means enabling individuals to have some control over the data related to their own health. This includes transparency about how the data is stored and used, and the ability to opt out of information sharing without it affecting the person's healthcare.

WHAT ARE THE NHS GUIDELINES?

NHS-specific guidelines are intended to ensure that confidential information is handled responsibly and in accordance with data protection laws.

These guidelines specify that every organisation providing health and care services must make every effort to:

- Keep data secure
- Use data that cannot identify individual patients whenever possible
- Use data to benefit health and care
- Not use data for marketing or insurance purposes (unless this is requested by the patient)
- Make it clear why and how data is being used

Opting out of large datasets

There are ways to opt out of having most types of your health data used for most purposes beyond your individual care (i.e. secondary/indirect purposes).

In the UK there are national variations in how this can be done. You can find out more about this, and how it can be done, here and in the [Further Information](#) section of this guide.

WHAT IS THE FUTURE FOR LARGE DATASETS?

The use of machine learning to improve healthcare is one possible benefit that large health datasets could help bring about. The scale of the NHS, and the amount of data that it stores, means that the UK might be particularly well placed to use this technology in the future. Some studies have already taken place that show its potential.

Machine learning is the process by which a computer analyses the information in a large dataset and detects patterns in it.

Machine learning in healthcare can enable computers to perform key tasks such as interpreting a clinical scan (for example, an X-ray or CT scan) or allocating beds in a ward with greater efficiency and speed. Developing and improving systems that can perform these functions requires the analysis of health data, from both healthy people and those with health conditions.

An example of machine learning in healthcare¹

Long stays in hospital are linked to worse outcomes for patients. However, it is not always easy to know which patients are likely to stay in hospital for a long time.

Researchers at one hospital trust in the UK set out to create a machine learning system that could predict which patients were most likely to become 'long stayers' (staying in hospital for more than 21 days).

The researchers used data from more than 1 million patient admissions to a hospital. Learning from this rich dataset, their system was able to predict which patients were likely to stay in hospital for a long time, with a high degree of accuracy.

In future, it is hoped that a system such as this could help tailor individual patients' healthcare, reduce the length of hospital stays, and improve outcomes for everyone.

¹. NHS Transformation Directorate: Using machine learning to identify patients at risk of long term hospital stays. 19 August 2021. Available at: <https://www.nhs.uk/ai-lab/explore-all-resources/develop-ai/using-machine-learning-to-identify-patients-at-risk-of-long-term-hospital-stays> (accessed April 2022).

WHERE CAN I FIND MORE INFORMATION ABOUT HEALTH DATA?

Information on specific subjects

NHS: Your NHS data matters

<https://www.nhs.uk/your-nhs-data-matters>

...a webpage that shows how records are shared for research and planning purposes.

NHS: Opt out of sharing your health records

<https://www.nhs.uk/using-the-nhs/about-the-nhs/opt-out-of-sharing-your-health-records/>

...a webpage where you can find out if, and how, you can opt out of NHS data-sharing.

NHS Digital: Data dashboards

<https://digital.nhs.uk/dashboards>

...a section of the NHS Digital website where you can explore some examples of the use of data in the NHS. You can also view datasets that have been de-identified and aggregated so that the data is anonymous.

Information Commissioner's Office:

Introduction to anonymisation

<https://ico.org.uk/media/about-the-ico/consultations/2619862/anonymisation-intro-and-first-chapter.pdf>

...a draft guide to the anonymisation of data, produced in 2021.

UK Government: UK National Data Strategy

<https://www.gov.uk/government/publications/uk-national-data-strategy>

...a national strategy that aims to drive the UK's data economy while ensuring public trust in data use.



WHERE CAN I FIND MORE INFORMATION ABOUT HEALTH DATA?

UK Government: Better, broader, safer: using health data for research and analysis

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

...a publication, authored by Professor Ben Goldacre, exploring health data in England and how its efficient and safe use for research and analysis could benefit patients and the healthcare sector.

UK Government: Data saves lives: reshaping health and social care with data

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

...a strategy for the future of data in health and care in England, and how high standards of privacy and ethics can be maintained.

Care Quality Commission: Anonymising data: Key questions for consideration

<https://www.cqc.org.uk/sites/default/files/Anonymisation%20Guidance.pdf>

...a detailed guide to the anonymisation of data, structured around important questions on this topic.

The what & why of trusted research environments

<https://understandingpatientdata.org.uk/news/what-why-trusted-research-environments>

...an article discussing trusted research environments – a type of secure data environment – the benefits they offer and the challenges they can present.

UK Data Service: What is the Five Safes framework?

<https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/>

...an explanation of the 'Five Safes' principles for data handling and how they can help protect data.



WHERE CAN I FIND MORE INFORMATION ABOUT HEALTH DATA?

Organisations involved in health data

Association of Medical Research Charities

<https://www.amrc.org.uk>

...an organisation supporting medical research charities in carrying out research that saves and improves people's lives.

Health Data Research UK

<https://www.hdruc.ac.uk>

...a national institute that aims to pull together the UK's health data to enable discoveries that improve people's lives.

National Institute for Health and Care Research

<https://www.nihr.ac.uk>

...a national institute that aims to improve the health and wealth of the UK by funding and supporting research that tackles health and social care challenges.

NHS Digital

<https://digital.nhs.uk>

...an organisation that manages the national IT and data services involved in patient care and health research.

Open Safely

<https://www.opensafely.org>

...a platform that allows for secure and transparent analysis of electronic health record data in a secure data environment.

use MY data

<https://www.usemydata.org>

...a UK group led by patients, relatives and carers that focuses on the beneficial and appropriate use of health data.

UK Health Data Research Alliance

<https://ukhealthdata.org>

...an independent group of healthcare and research organisations that aims to establish best practice for the ethical use of UK health data for research.

Understanding Patient Data aims to make the way patient data is used more visible, understandable and trustworthy, for patients, the public and health professionals. <https://understandingpatientdata.org.uk/>

© 2022. This work is licensed under a CC BY 4.0 license.

