A review commissioned by the Secretary of State for Health and Social Care

Better, Broader, Safer: Using Health Data for Research and Analysis

Summary / April 2022

Contents

Goldacre, B & Morley, J. (2022). Better, Broader, Safer: Using health data for research and analysis. A review commissioned by the Secretary of State for Health and Social Care. Department of Health and Social Care.

Ministerial introduction

Data has transformed our world in powerful ways. It can connect us, help us make better decisions, and enable life-changing discoveries. In every field, from agriculture to finance, things that once seemed impossible have become commonplace.

In some ways, health data is unlike other data. Concerns about privacy take on an even bigger life when it concerns our personal medical data. Moreover, the systems across the NHS and medical research can feel intimidatingly complex. Yet in other ways, healthcare is more suited to data and the innovation that follows than almost any other sector — with the depth and coverage of NHS data providing unique opportunities. Navigating complexity can come with even greater gains, and the number of applications for medical data in health research are seemingly never-ending. The rewards of getting it right are profound, with not just lives saved but longer, healthier and happier lives too.

There's no better proof of this than how we embraced data to respond to the pandemic. Even with Covid ongoing it was vital we did all we could to capture the gains we'd made, so last year the government commissioned Ben Goldacre to deliver this report into the use of health data for research and analysis. I'm grateful to him and his team for this work. He has certainly met our level of ambition with some 185 wide-ranging recommendations for us to explore.

This report shows that we need to be as thoughtful as we are innovative, guided by safe ethical frameworks for providing access to data, as well as systems that ensure under-represented groups are well represented. It also makes clear that we have all the building blocks we need for success, including an unrivalled wealth of experience in using health data. However, it also shows areas where we must boost our capability and capacity if we are to reach our full potential.

Soon we will be publishing the final version of our data strategy, Data Saves Lives, which will set out how we will unleash the enormous potential of data in health and care. It will include our response to these recommendations, many of which have already helped to shape our work in digital transformation. For example, we have already announced up to £200 million to invest in the development of Trusted Research Environments and digitally enabled clinical trials.

If we put this agenda into action, then I am confident that the future of health research will be bright, and that data will drive the longer, happier and healthier lives that we all deserve.

Sajid Javid
Secretary of State for Health and Social Care

Foreword

The NHS has some of the most powerful health data in the world. Almost every interaction with the health service leaves a digital trace: the diagnoses, treatments, tests and outcomes for almost every citizen in the country.

This raw information has phenomenal potential. Data can drive research. It can be used to discover which treatments work best, in which patients, and which have side effects. It can be used to help monitor and improve the quality, safety and efficiency of health services. It can be used to drive innovation across the life sciences sector.

But raw data is not powerful on its own. It must be shaped, checked, and curated into shape. It must be housed, and managed securely. It must be analysed. And then it must be communicated, and acted upon. That work all requires people, with modern data skills, in teams, using platforms that protect patients' privacy and avoid needless duplication of effort.

This review sets out a practical vision of how we can collectively achieve this goal.

We are pleased that some of our early recommendations have already resulted in action, and particularly encouraged by the recent announcement of £200m for Trusted Research Environments. Building these platforms will be challenging. But it can be done by starting small, meeting common use-cases first, and building strong teams.

On behalf of the team I am deeply grateful to the many people who have enabled us to see so far into the system and its needs, including Ministers and staff at the Department of Health and the NHS. We are particularly grateful to the team at NHSx, now NHS England, who supported our work throughout. Our Senior Stakeholder Group gave excellent advice to keep our work firmly on target.

More than anything it was a fascinating and rare privilege to be able to discuss health data in detail with over 300 people in individual and small group discussions; and a further 160 people in a series of single sector focus groups.

We have set out to repay this generosity by being clear. The full review text is long, and contains substantial technical detail. This is for good reason: the challenges themselves are technical, and this reality can never be wished away.

But there is every reason for optimism. Modern open working methods can avoid duplicated effort, and drive efficient delivery. The NHS has already collected unparalleled lifetimes of data, from tens of millions of patients, in thousands of organisations, over endless decades of effort. Secure platforms can be built for less than the cost of digitising one hospital. If this job is done well, then the system can finally unleash the full power of all NHS data ever collected, in one fell swoop.

Professor Ben Goldacre
April 2022

Summary

Introduction

The Power of Health Data

Data is at the core of all good work in healthcare. Data is how researchers and clinicians learn which treatments work best, which interventions have side effects, and which patients are most likely to experience them. Data can help innovators to evaluate existing treatments, and see what works best for which patients; but it can also help them develop entirely new kinds of medical technology. Data can help analysts find new opportunities to improve the quality, safety, and cost effectiveness of care, by monitoring and evaluating all health service activity and outcomes, for all patients, and all clinicians, in all organisations, across the whole of the NHS. Where problems are found, data can be used to target new training interventions, or simple feedback to improve care, or even roll out wholesale re-design of local services: after doing so, data helps analysts track whether those interventions were successful at changing activity and outcomes. In the hands of academic researchers data can be used to identify fundamental truths about the risks of environmental exposures, or the reversible causes of death and disease; in collaboration with the life sciences sector data can be used to refine medications, or develop whole new classes of medical intervention.

Data does not do this on its own. It needs to be managed, reshaped, prepared, curated, and cleaned by people, and well-designed systems combining humans and technology, using software and platforms that facilitate sharing and re-use of prior work. Then that data needs to be analysed. All this work is done thoughtfully by people in teams, who need to exhibit a vast range of skills, whether combined in individuals or collectively as a group. This includes generalist data science skills such as data management, statistics, software development, technical documentation, and data visualisation. But it also includes more specific skills and knowledge related to the domain of healthcare: how health services operate; what the codes and data for blood tests, treatments and diagnoses mean in the real world; and how that information is stored in the everyday Electronic Health Records (EHR) of real patients.

Context

This review was initiated with a very broad Terms of Reference, covering a wide range of challenges for better use of data in England (see Appendix) with the patient at the heart of all good work. Fundamentally these challenges all reduce to one simple question: how can we get better, broader, safer use of NHS patient records, to drive innovation and save lives?

The answers are somewhat technical, and this should not be shied away from. It is easy to say that NHS data is powerful, and that we must listen to patients. Acting on these impulses to deliver change is a very different business: it requires some technical understanding and insight. Because of this, the review is presented at three levels of detail. The Executive Summary contains a short overview of high level strategic opportunities. This Brief Summary contains a longer overview of the opportunities, challenges,

and recommendations. The Full Text contains a detailed explanation of the work, with extensive findings from interviews and desk research; alongside practical descriptions of the mechanics of working with NHS data, to help ensure informed discussion; and granular detail on the practical and technical aspects of recommendations, where these are needed. The objective is that all can participate in an informed discussion around the best working methods, and ensure strong forward progress. NHS data is a challenging space, with huge opportunities, but modernisation of the workforce, working practices, and platforms is long overdue: it will only come when the system engages robustly, at the right level of technical detail.

Two forces of history have defined the context for this review. The first is COVID-19. The pandemic has shown more than ever the need for fast flowing, detailed data on a huge number of patients to manage the nation's health effectively. To illustrate this collective challenge: the first wave of the pandemic came, and then went away entirely, before a single COVID-19 case appeared in the conventional, slow-flowing NHS data research extracts; and there is still no way to see, at a national level, the identity of each patient admitted to hospital with COVID-19, linked onto their vaccination status, medical history, and other information that would help inform urgent work on changes in vaccine effectiveness over time, or the extent to which new variants are covered.

The second force is more technical, but nonetheless historic. During 2021 the NHS attempted to implement an ambitious programme extracting and aggregating the coded GP records of every patient in the country (excepting those who have opted out), with identifiers such as name and address removed, in order to then disseminate this data out to multiple users for various health and social care purposes. This project was suspended in July amid widespread privacy concerns from professionals and the public, and an estimated 1.5 million patients opted out of their records being accessible for planning and research. This followed the pattern of the Care.Data programme in 2013, but with one crucial, positive difference: the work was successfully re-railed. This was achieved with a clear public commitment that detailed GP data at national scale would only ever be accessible inside a Trusted Research Environment, where data misuse can be obstructed and detected, and where every action on the data can be publicly disclosed to earn public trust.

A Platform Approach

The significance and importance of this should not be under-estimated: it represents a fundamental, positive, and long-overdue sea change in the way that NHS records are stored and used. Collectively, NHS records represent a dataset of unprecedented depth and breadth in the history of humanity. They have been collected over many decades, covering the entire medical history of tens of millions of patients. Collecting this data has been a phenomenal enterprise across the whole NHS. The system is within a hair's breadth of being able to capitalise on this huge investment for the purpose of saving lives, unlocking the power that lies within the data. This can be achieved, but only by engaging robustly, practically, and technically with three core needs: the need to build secure platforms for analytics; the need to build a skilled, technical workforce with software skills; and the need to embrace modern, collaborative approaches to computational data science, where all code is shared as an open resource for re-use by all.

This can all be done, through small and large steps. The full reasoning and practical recommendations are contained in the pages that follow. This review was conducted over a period of six months, with extensive interviews and equally extensive desk research: the team spoke with over 300 individuals one on one, or in small groups, conducted 8 open focus groups, and received over 100 written submissions. It has been an exhausting honour - and a fascinating pleasure - to see the system so close up, and from so many different perspectives.

The challenge of privacy in health data

Managing a health service effectively, or delivering high quality research at national scale, requires that analysts have access to the most detailed information, across the health records of every individual in the country, to do their good work. This also means a growing group of trusted analysts having access to every recorded detail, of every medical event, for almost every citizen, all the way back to birth. Patients, professional groups and campaigners are rightly concerned about patients' privacy being protected when large volumes of data are accessed for analysis, research and innovation. Managing this problem - widening access to records, while also preserving patients' privacy - is the fundamental challenge for use of NHS data in service improvement, academic research, and the life sciences sector. The NHS must maintain trust and active enthusiasm from patients and the public. Researchers and analysts, conversely, are deeply frustrated by innaccessibility of data, and missed opportunities to improve patient care, when slow information governance processes obstruct data access.

Pseudonymisation and contracts

It is important that the system recognises the challenges in current approaches, to have a pragmatic discussion about better working practices. At present the NHS relies excessively on two techniques to protect privacy: pseudonymisation; and trust in individuals and organisations, administered through contracts.

Pseudonymisation is the process of removing "direct identifiers" such as name, date of birth, and address from records before sharing them to a wide pool of users. Where pseudonymisation is combined with other organisational and technical controls it can be somewhat helpful; but it is common to find examples of its benefits being overstated, or relied upon excessively. In reality, pseudonymisation is easily reversed when working with very detailed data such as NHS patient records.

Knowing the approximate date range in which someone had a medical intervention, their approximate age, and their approximate location is often enough to re-identify someone in a pseudonymised dataset, and then - illegally - to see everything else in their record. Women face particular concerns: knowing someone's approximate age, approximate location, and the approximate time at which they had children can also often be enough to make a confident unique match; this is the kind of information that will be known by someone at the schoolgate, or a colleague. This is not to say that health data users are untrustworthy: but the system must be resilient to untrustworthy users; and it is well documented that other large administrative national datasets are sometimes misused.

Importantly, the risk of re-identification in pseudonymised data increases as the dataset grows to cover a larger proportion of the total population, and as datasets become more detailed. This has important implications for all plans to gather large volumes of detailed data about the whole population, such as the GP Data for Planning and Research programme. Furthermore, when the number of people accessing a dataset grows, there is an increase in the small risk of there being untrustworthy individuals among those with access. This is important. The vast majority of those accessing data are trustworthy and abide by the law. However it is important not to downplay risks: there are many examples - in medicine and in other sectors - of some people misusing large datasets to which they have access.

Because of the security shortcomings inherent in widespread dissemination of pseudonymised data, the system has additionally needed to rely on contracts and trust, administered through complex regulatory frameworks and systems to decide who can have what data. This approach brings two problems. Firstly, it creates very substantial anxiety for individuals giving permission for each data dissemination: this makes the system cautious, and slow, creating deep frustration (and many abandoned projects) for analysts, researchers, and innovators. Secondly, this approach will always inherently struggle to scale to larger numbers of users, which is a key ambition for better use of NHS data. Pseudonymisation, alongside trust and contracts, has also not been sufficient on its own to fully reassure patients and professionals.

Privacy concerns and public support for use of data

Privacy concerns are at the heart of objections to large scale NHS data sharing projects from professionals, campaigners and patients. These concerns have derailed large NHS data projects on two occasions: the 2013 care.data programme, and the recent initial planned work on the GP Data for Planning and Research. Both of these projects aimed to collect significant amounts of the clinically coded data captured in the GP records of every citizen, and then disseminate varying amounts of data on, in pseudonymised form, to various NHS and external users, after an application and approval process. Both projects resulted in large scale concern from patients and professionals. Both resulted separately in very large numbers of patients opting out of their records ever being shared outside of their GP practice (approximately three million by the end of 2021) with opt-outs now at a scale that will compromise the usefulness of the data. It is crucial that the shortcomings of pseudonymisation are not downplayed or ignored. Wherever this is done, it undermines public trust and causes conflict between the NHS and the professional groups, campaigners and patients concerned about patients' privacy. It is important to communicate and advocate to the public about the power of NHS data, but ultimately trust is earned by the system taking provable, credible steps to protect patient privacy, and by being transparent with everyone about everything that is done with their deepest medical secrets.

The future

Fortunately there is a clear path forwards. In many other sectors - such as census work at ONS, for two decades - data is not disseminated out to users. Instead the analysts go to the data, and work inside a secure platform called a Trusted Research Environment. This working style must be adopted in the NHS.

The recent announcement that the GP Data for Planning and Research dataset will only be available in a Trusted Research Environment is therefore extremely welcome. It is clear that a robust TRE meets the privacy concerns expressed by the community, and will facilitate a smooth transition to the NHS having greater access to data. It is crucial that this policy stance is maintained. There is no new privacy emergency, but further expanding the population coverage and granularity of data aggregation and expanding the pool of data users aggregation and dissemination of data should not happen until TREs are in place. It is crucial that all data access happens in platforms where any potential misuse is obstructed, and easily detected. As a general principle - while the current legal arrangements around pseudonymised data seem to be overall unclear - all pseudonymised national detailed health datasets that are vulnerable to re-identification with additional information about the individuals included should be treated similarly to those that have name and address in the clear, both practically and in governance, regulatory, and legislative frameworks.

There is additional important context for this choice, and the GP data for Planning and Research. At present, the system as a whole tends to only discuss, and see, the uses of data at the centre, in national organisations such as NHS Digital. However, due to the absence of secure analytics platforms, and as a consequence of each single GP practice and NHS Trust acting as an independent data controller, there is now a large, poorly documented, and poorly understood network of data disseminations out of local organisations. In particular, large volumes of GP records are regularly exported to multiple other systems for analysis, research, and other activities, often in off-site environments containing many hundreds of practices' patient data. These exports are approved by individual GP practices, creating a substantial time burden and responsibility for clinicians in evaluating each extract, and this in turn creates other unintended consequences. For example, it is common to find that a practice has approved some general purpose research data flows, but not others: it is unclear whether this reflects a deliberate decision, or a combination of happenstance and the persistence of requests. The ambition for a single GP data extraction aims to help resolve this situation by replacing these myriad disseminations with one single system, improving oversight, and reducing the burden on GPs to evaluate multiple complex requests for bulk data. National GP data flowing into a TRE is therefore an important privacy safeguard for patients, a substantial net improvement in protections for patients, and a reduction in burden around data flows for GPs.

The full text of the review also considers other forms of risk mitigation including: removal of "sensitive codes" (which obstructs research on key areas of medicine); data minimisation (which has uses but is under-researched); sub-sampling (which has limits when aiming to detect subtle statistical signals); data perturbation (which has a role but requires a substantial research programme, and is complex to implement); and emergent methods such as "homomorphic encryption" (which has seen no substantial working health implementation to date). Overall they show that this an important area of work which has been relatively neglected. Wider access to NHS patient records requires that the system as a whole takes the challenge of practical approaches to secure analytics, developing and evaluating robust methods for protecting patients privacy at scale. There is a clear role for UKRI/NIHR in providing open, competitive resource for applied methods research into privacy preservation, to earn public trust, in collaboration across the NHS, epidemiology and security engineering communities. By building great platforms, we can harness the untapped power in all NHS data.

Trusted and Shared Research Environments

The current paradigm of disseminating extracts of data out to multiple different locations creates very substantial problems, well beyond the security challenge. It duplicates risk, by housing sensitive data in multiple locations, with limited central oversight; but it also duplicates cost, by creating multiple different technical implementations and governance arrangements. It reinforces monopolies around data access, by creating complex unseen powerbases around datasets; and it duplicates effort, by obstructing re-use of code for curation or other common tasks. This in turn also reduces analytic quality, and efficiency.

Moving to working with NHS data in shared TREs will address all these challenges. Analysts, researchers and innovators can come to the data, and work on it securely, in situ, without downloading it off site, using standard environments that share code and working practices. This will improve access, but also data quality and efficiency, allowing all new users to benefit from the curation and analysis work of all previous users, in settings that have strong technical documentation and clear working practices.

This should be recognised as a large job, but absolutely crucial. It will protect patients' privacy; permit reform of obstructive IG rules created to manage less secure and outdated options; facilitate substantially wider access to data; facilitate modern open working methods; and create a rapid explosion in the efficiency, openness, and quality of analytic work. This approach is also strongly supported by the Life Sciences Vision from the Office for Life Sciences. Previous reviews and strategies, most notably the Tech Vision (2018) and Personalised Health and Care 2020 (2014), promised to ensure NHS data was stored in a single secure location, but did not identify the means for achieving this goal. Instead of a single access location, this work therefore created a data collection and dissemination function (NHS Digital) sending data out to multiple other locations for use. TREs are the correct answer to this challenge.

Strategy

The system should be cautious around imagining that it can push away the challenge of TREs - and all work with NHS data - by procuring "black box" services. Building platforms, capacity and modern working methods for data is a complex technical challenge, requiring deep knowledge across a range of domains: data science, data architecture, and software development; but also clinical informatics, NHS data needs, health data research, and more. This work must be done close up with real users of data, constantly iterating to improve platforms and approaches. There is no single contract that can pass over responsibility for this work. These new and complex technical challenges around data must be met by building teams, tools, methods, working practices, code and platforms.

A TRE should be conceived of as having three components: a service wrapper; the underlying generic computational and database services; and the bespoke software needed for work with NHS data. The service wrapper should be a common framework used by all TREs to implement permissions for projects and analysts, check that outputs are non-disclosive, publish activity logs, and achieve other similar tasks: there is no merit in the current duplication and inconsistency currently seen for this work. The compute and database aspects of a TRE are largely generic tasks that can be readily delivered by staff with strong generalist software and data science skills: this is important, as such staff are more easily recruited from other sectors.

The challenge of creating bespoke code specific to the needs of NHS data management and analysis will require the system to foster an open collaborative ecosystem, creating code and methods as described in the sections below on Modern, Open Working Practices for NHS data. This is a normal challenge for any community of data users to address: outside of commonplace data needs, such as those in accountancy, it is routine for analysts and communities to meet the challenge of developing bespoke code and working methods for their bespoke needs. The additional challenge for working with NHS data is that the user community is so large and diffuse: this necessitates an open and shared approach to all code and technical documentation. Developing these shared methods, tools, code and working practices will require a mixture of open competitive funding from funders and the NHS, for innovation in NHS data management and analysis; and national strategic work to surface prior art hidden in local teams.

Recent policy commitments for the new national GP data extract to be "TRE only", and to build a national TRE for this work, are very welcome and should be built upon. Other national datasets such as SUS/HES are smaller, less detailed, and can therefore be accommodated alongside GP data in a TRE at minimal marginal effort. All large, detailed, disclosive national datasets should in the future only be available in a national TRE, even when they are pseudonymised; however where patients have actively consented for their data to be sent to other data centres (for consented clinical trials or research studies) this should be respected.

What to build

To meet these needs there should be no more than three national TREs. It is helpful to build more than one national TRE to address two key risks: monopolies around access; and the risk of non-delivery, or poor service. Every TRE containing national NHS data should be a shared resource where all NHS and other users can apply for access: whenever a "TRE" is run as a closed service for internal use in only one organisation, it drifts away from the open working methods and robust service wrapper needed to earn public trust and deliver high quality analytics. All TREs should support and require modern, open approaches to data science, as set out in the section on Reproducible Analytic Pipelines below.

Alongside national TREs there will be circumstances where smaller satellite TREs are necessary, although these should be minimised where possible. Integrated Care Systems are new organisations in the NHS, all using data to improve the quality, safety and efficiency of care. The closed, duplicative work of the past on local data analysis environments operating as "black box" services should not be repeated. All local TREs for ICSs should ideally conform to a single national model, with pragmatic flexibility to account for diverse local datasets. Procurement should be focused on the methods, code, tools, and approaches that can be used in all TREs; not for whole TREs as a single closed unit as seen in the past. All local TREs should support and require modern, open approaches to data science.

Alongside local NHS TREs there are two further categories of data that require great security, accessibility, and usage. Firstly, the national audit, registry, and quality improvement projects (which are separately overdue a strategic review): a very large number of bespoke data collections and NHS data flows used to monitor and improve services, or conduct subject-specific research. These are often "labours of love", with inspiring and committed teams, but are generally treated as isolated, standalone datasets, when many would be better implemented as thriving analytic communities inside a shared data resource. Secondly, there are numerous bespoke research data collections, such as the birth cohorts, and other diverse datasets. Here there is a need for caution: some senior leaders expressed concern that platform work here has historically been conducted and managed behind closed doors, with unclear delivery. Despite this, for both national audits, and research datasets such as cohorts, there are several very strong examples of mature, ambitious teams ready to adopt TRE working and modern open methods.

To make change practical, the best route forward is to identify pioneers in each of these settings who are most ready to fully embrace open methods and TRE working, to light the way for others: three ICSs; three national quality improvement registry or audit teams; three academic birth cohort or electronic health record analysis teams; and 1-3 national NHS analytic teams. These should be selected competitively as those with the best current technical skills. This can be in parallel to "business as usual" in their organisation, but should incrementally subsume it.

It is crucial that TRE work is modular, developing methods, working practices, code and tools that are shared across all TREs, rather than procuring closed "black box" services as in the past. There are many single tasks that UKRI/NIHR could usefully fund work on. The list of examples below is not provided as a comprehensive or prioritised programme of work, but rather as an illustrative list of the kinds of work, some reflecting ongoing activity at various universities, that funders could usefully support through open competitive funding. Work of this kind will help to drive a rich, competitive and collaborative ecosystem of code and methodological approaches to meet key emergent challenges and support work by all.

- **Methodological innovation and code for Data Curation**, developing best methods to make complex NHS data ready for efficient high quality analysis, as discussed in the next section.
- **Methodological innovation and code for data minimisation**, reflecting the fact that minimisation is commonly used to protect patients' privacy, but with little formal or quantitative guidance for decision-makers to determine the correct amount of information to release about each individual in a dataset. Applied methodological work and code tools in this space would meet their needs, draw on deep theoretical work around disclosure and privacy engineering; deep domain knowledge around clinical records; information governance requirements; and similar.
- **Methodological innovation and code for detection of data misuse**, reflecting the fact that data analysis environments commonly keep logs, but these are currently under-used, or only examined manually. To meet the strong desire for wider access to innovate in NHS data, there is a need for more robust and scalable approaches to monitoring users' activity, drawing on various deep technical domains and skills.
- **Methodological innovation and code to detect unwarranted variation in care**, meeting a key common analytic need for NHS service analysts. There is extensive prior art in this space, and numerous challenges such as avoid over-inclusive or insufficiently sensitive algorithms; with huge scope to take this prior art, evaluate it, and scale it across the NHS in national and local TREs.

- **Methodological innovation and code for federated analytics**, reflecting that this headline challenge overlies a complex set of tasks drawing on deep technical skills around research software engineering, but also very deep technical domain knowledge around health data analysis, and the kinds of data to be accessed and curated, with different approaches needed for different forms of meta-analysis combining different intermediate elements from single data centres, and similar. These complex and entwined challenges around methods and implementation will not be met well by delivering "federated analytics" as a closed black-box service.

Work of this kind will help deliver usable data, in performant platforms, for use by NHS analysts, researchers, and the life sciences sector. To maintain focus on delivery, TRE work should be coordinated and executed by teams or institutions with a sole focus on only providing platforms to help other people achieve their analytic tasks. Funding for methods and code around elements such as curation and secure analytics should be open and competitive, to ensure the best ideas and teams are identified and amplified.

Summary

This work is readily deliverable. If it is done, the UK will have well-curated national and local data, with shared code that makes projects fast to initiate, complete, and spread. It will deliver enhanced security and transparency, making it safe for the NHS to grant data access to a wider pool of individuals and organisations. It will permit development of a "fast track" through the current onerous IG requirements, reflecting the lower risks presented by TRE access. It will make NHS statistics and research outputs more trustworthy and reliable, by facilitating Reproducible Analytic Pipelines and modern open working methods as the default. Overall it will drive research, innovation in life sciences, and better use of data to improve the quality, safety and efficiency of NHS services.

The full text and recommendations contain detailed background, alongside detailed practical recommendations on how TREs can be rapidly developed to meet user needs, their core characteristics, and methods to work around organisational and technical barriers to delivery. The summary recommendations below link out to these more detailed recommendations.

TRE Recommendations

1. Build trust by taking concrete action on privacy and transparency: trust cannot be earned through communications and public engagement alone.

2. Ensure all NHS data policies actively acknowledge the shortcomings of "pseudonymisation" and "trust" as techniques to manage patient privacy: these outdated techniques cannot scale to support more users (academics, NHS analysts, and innovators) using ever more comprehensive patient data to save lives.

3. Build a small number of secure analytics platforms - shared "Trusted Research Environments" - then make these the norm for all analysis of NHS patient records data by academics, NHS analysts, and innovators, wherever there is any privacy risk to patients, unless those patients have consented to their data flowing elsewhere. Every new TRE brings a risk of duplicated effort, duplicated information governance, duplicated privacy risks, monopolies on access or task, and obstructive divergence around data curation and similar activity: there should be as few TREs as possible, with a strong culture of openness and re-use around all code and platforms.

   Detailed recommendations on establishing national TREs are in TRE 1-9, TRE 23, TRE 53-55; standardising the approach to local NHS data platforms TRE 24-36; ensuring delivery of performant accessible shared TREs for academic research TRE 40; academic TREs should use standard NHS approaches where available TRE 41, 42; consider common TRE infrastructure TRE 43; funding and amplifying skilled teams for TRE work through open competition, coordinated by people with data architecture skills TRE 46-51; detailed recommendations on avoiding short-term or closed funding, that props up legacy working Open 42, TRE 50, Open 33, Open 35, Open 37; funding TRE and software projects distinctly from academic research papers TRE 51, Open 34, Open 39, and Cur 15; detailed recommendations on academic TRE funding TRE 55; academics using NHS TREs to access NHS data TRE 40; the need to fund AI TRE work separately TRE 57.

4. Use the enhanced privacy protections of Trusted Research Environments (TREs) to create new, faster access rules and processes for safe users of NHS data; ensure all TREs publish logs of all activity, to build public trust.

   Detailed recommendations on standard governance and transparency are in TRE 11-17; detailed recommendations on making data access faster after secure TREs are implemented can be found in IG 9-11 and 13-15.

5. Map all current bulk flows of pseudonymised NHS GP data; then shut these down, wherever possible, as soon as TREs for GP data meet all reasonable user needs.

   Detailed recommendations to help identify and disclose existing data flows are in TRE 16-17; using TREs to replace existing data flows TRE 18, 21-22, 38 and 56; maintaining public trust in TREs TRE 19-20.

6. Use TREs - where all analysts work in a standard environment - as a strategic opportunity to drive modern, efficient, open, collaborative approaches to data science.

   Detailed recommendations on designing TREs to support modern open working are in TRE 10, 39, 44, Open 42, 45; using TREs to achieve culture change TRE 37, 45, and 52.

Modern, Open Working Methods for NHS Data Analysis

Raw data - such as NHS patients' electronic health records - is prepared, analysed, and visualised by writing code that issues instructions to computers. Data preparation and analysis are hugely complex technical tasks. This work is not done by isolated individuals, but rather in huge arcing chains of mutual interdependency, writing complex code across multiple teams and organisations.

Modern methods to manage complex technical work

There are well established methods for imposing systematic order on this kind of challenging complexity in other settings: developing code interactively, and collaboratively, in industry-standard systems that allow teams to track, annotate, and attribute all changes; writing adequate technical documentation that sits alongside the code for all subsequent users or viewers; taking recurring tasks and turning them into "functions" that are regularly re-used; and so on.

At present too much work with NHS data, at all steps of curation and analysis, in all sectors, is done behind closed doors, often driven by thoughtless defaults rather than any strong motivated decision to support closed working. The review team was given multiple clear examples of situations where code or methods used to create insights for service analytics, or research, were actively withheld; in ways that held back replication, critical review, validation, implementation, re-use or improvement of the work; and seemed to serve no clear strategic national benefit.

The Office of National Statistics and the Government Digital Service have already developed, over recent years, a set of best practice principles for modern, open, collaborative work with data. This work is branded as "Reproducible Analytical Pipelines" (RAP) with a clear set of design principles to support high quality analytics that are reproducible, re-usable, auditable, efficient, high quality, and more likely to be free from error. At minimum a RAP will meet various criteria. It will minimise manual steps (such as copy-paste, point-click or drag-drop operations; where it is necessary to include them, they must be properly documented). It will be built using open source software for data management, analysis and visualisation (such as R or python) as this is standard, portable, and available to all for checking and re-use. The code will be open to anyone for review and re-use, with all code shared openly through open standard file and code sharing platforms such as GitHub. The code will be well "commented" with adequate documentation embedded within the work. These working practices, alongside good practice for code review and quality assurance, improve the quality and efficiency of work with data.

Adopting modern open methods for NHS and academic data analysis

The RAP community coordinated by ONS across multiple government departments has extensive experience of training and culture change: this should be drawn upon. The NHS analyst community could make this transition swiftly, not least as part of a long overdue modernisation of career structures and working practices, as discussed in the section below on supporting and modernising that professional group. Due consideration must be given to the broad range of tasks and skills in the NHS analyst profession: from those doing technical data preparation and analysis (who should use RAP); through to those who specialise in tasks such as data communication (who should work alongside those using RAP).

The academic research community working with NHS health data faces some different challenges: it is world class at delivering conventional individual research paper analyses, due to the inherent richness of NHS data, and the success of open competitive research funding in this space. However, for foundational work such as data curation, secure analytics, and efficient open computational working there is almost no open competitive funding, little recognition, and therefore poor progress. More concerningly, the the review team were given examples of funding for these foundational and platform tasks being diverted onto traditional academic research paper analyses in single clinical topics, which have historically been regarded - unhelpfully - as having unique and higher status. This is problematic, as the foundational work is key. A focus on methodological innovation and open code for core tasks can deliver an explosion of outputs across all data users, dramatically reduce the startup time for each analysis, and facilitate strong technical collaboration between NHS analysts, academia and the life sciences sector, built around a culture of shared code and technical documentation with low entry barriers, rather than meetings.

As a related issue, the academic community working with health data has also been slow to recruit, recognise, or use the skills of software developers appropriately. Conversely, progress on this has been strong in adjacent academic fields such as structural genomics, physics, or structural biology, where there is a longer and deeper tradition of sharing code, and sharing credit with expert software developers. Again this is a function of context and history, rather than good will: any strategic transition to involve developers in academic work with health data will require support from universities and funders, not action from individuals. As very positive context, the Research Software Engineers community has grown rapidly over the past decade in the UK, developing and sharing applied practical skills to work alongside researchers as equal collaborators on novel and creative academic output. The RSE community should be energetically supported to expand its work into health data.

Addressing myths about open working

Because open working is somewhat new to some in the health data space, it is important to address some myths or possible misunderstandings. Adopting open working practices does not mean other countries or industry can exploit intellectual property created with state funds: there should be a robust and thoughtful exceptions framework to impose commercial licenses or restrictions on review and (separately) re-use of publicly funded code, where this is actively helpful; but this closed approach should be used in a planned and deliberate fashion, where it meets national strategic objectives, not as the unplanned default approach. Code, methods, tools and documentation for well curated data and performant analytics platforms should be regarded as a national asset that will draw investment and drive productivity: not something to have hidden in closed "black box" services and teams.

Related to this, open working is fully compatible with use of commercial products: it requires only that new code and methods created for and funded by the state should be shared as default, for interoperability, quality, and efficiency. Similarly, open working does not mean that noboby is paid: simply that new code and methods are contracted from the outset as a buy-out; during interviews there was strong support - including from contractors - for this approach.

In addition, open working does not mean that the results of every analysis must be shared openly, or in real time. The results of an analysis are separate to the code and methods used to create them. It may often be reasonable for NHS analysts to run data analyses to monitor and optimise the delivery of care, for example, without disclosing the results of all such analyses publicly in real time: organisations should be free to use data without always fearing distraction from "performance management through the media"; and the rights or wrongs of this are a separate discussion to the question of sharing code, methods, and technical documentation for analytic work.

Lastly, open sharing for code is not a philosophical, political, or ideological stance, but rather a practical one. Data curation and analysis is complex technical work across multiple teams, and it can only be done well where technical material (such as code, methods and documentation) is shared between those teams. In the commercial sector, this sometimes means sharing code privately among a small group of staff. But the people working on NHS data stretch across hundreds of diverse public and private sector organisations. Creating a closed permissions-based system to carefully police limited sharing among a huge array of individuals across all these organisations would be a vast technical and bureaucratic project, of inconceivable complexity and expense. Most importantly, this expensive approach to balancing closed working and accessibility of

information would bring no clear benefit, as there is no clearly articulated need for code and methods to be withheld from wider access.

Conclusions

By taking a platform approach - and adopting modern, open working methods - analytics with NHS data can transition from a dispersed community, with entry points based on meetings and relationships, into a rich, open, ecosystem where innovators from all sectors can efficiently identify opportunities to contribute and benefit.

The following high level recommendations will help achieve this goal. They map onto detailed recommendations, and background, in the full review text.

Recommendations

7. Promote and resource "Reproducible Analytical Pathways" (RAP, a set of best practices and training created in ONS) as the minimum standard for academic and NHS data analysis: this will produce high quality, shared, reviewable, re-usable, well-documented code for data curation and analysis; minimise inefficient duplication; avoid unverifiable "black box" analyses; and make each new analysis faster.

   Detailed recommendations in Open 1, 2, 14.

8. Ensure all code for data curation and analysis paid for by the state through academic funders and NHS procurement is shared openly, with appropriate technical documentation, to all data users. Data preparation, analysis and visualisation is complex technical work, requiring collaboration by many individuals, who may never meet, in a range of organisations, across the NHS and other sectors. The only way to manage this shared complexity is by sharing information, as in other technical fields.

   Detailed recommendations on the role of clear guidance and policy in supporting open code are available in Open 6-9; writing an Open Analytics Policy Open 14; open working in standard NHS analytics contracts Open 15; an exceptions framework Open 4; clear statements from regulators (Information Commissioner, MHRA, Health and Care Information Governance Panel) Open 10-12; produce clear guidance on disclosure risk and open code Open 46; the role of contracting and procurement in promoting modern open methods Open 3 and 15; negotiate co-ownership of claimed commercial innovations from NHS data Open 13, IG 24; Data Controllers should require RAP and open code sharing from data users Open 7; commission intermittent open code audits to drive improvement Open 16; research funders promoting open code through funding contracts Cur 4, Open 3, 6, 15, 29, 30; mechanisms for when publicly funded code is withheld Open 5; technical writing and documentation function Open 17; the role of TREs in promoting modern open approaches as a default Open 42, 43, TRE 10, 39, 44; TREs themselves should be built on principles of RAP and open code Open 43.

9. Recognise software development as a central feature of all good work with data. UKRI/ NIHR should provide open, competitive, high status, standalone funding for software projects and developers working on health data. Universities should embrace Research Software Engineering (RSE) as an intellectually and academically creative collaborative discipline, especially in health, with realistic salaries and recognition.

   Detailed recommendations on the role of universities in promoting the importance of software development for research are available in Open 21-28; the role of academic funders in promoting modern open methods Open 29-30, 33-40; working group to develop an attribution model for re-use of code and data Open 24; authorship for software developers and data scientists Open 25; address sharing during the COVID-19 pandemic Open 26; three pioneer Research Software Engineering groups in health data Open 28; open funding for health projects and programmes focused on code Open 33, 35 and TRE 49; treat data infrastructure as open code Open 34; review prior delivery of open code by applicants when considering funding for new code projects Open 36; ensure experts on code select and oversee code projects Open 37; ensure objectives and outputs of code investments are open Open 38; ensure funding for code and platforms is not diverted onto single topic academic papers Open 39; avoid "regressive funding models" built around short-term bursts of funding Open 40; sustainability for software projects Open 41; modify the Research Excellence Framework (REF) to reflect computational work and require code for data-driven research papers Open 21; build on work from Wellcome Data Science team on best practice in code for health Open 33; TRE work to resource TRE 55.

10. Bridge the gap between health research and software development: train academic researchers and NHS analysts in contemporary computational data science techniques, using RAP where appropriate; offer "onboarding" training for software developers and data scientists who are entering health services research and epidemiology; use in-person and online training; make online resources openly available where possible.

    Detailed recommendations are in Open 18-20 and 31; fellowships for software developers in health data Open 32.

11. Note that "open code" is different to "open data": it is reasonable for the NHS and government to do some analyses discreetly without sharing all results in real time.

Data Curation

"Data management" or "data preparation" is the crucial first step of any meaningful data analysis. The team has spoken to a large number of coalface NHS data analysts and researchers during the course of the Review: they overwhelmingly expressed frustration at the scale of duplicated effort in this space. The Association of the British Pharmaceutical Industry (ABPI) have said that they estimate 80% of all work on an analysis project using NHS data is spent on this data preparation, and they have previously recommended that 80% of the national resource deployed on data science in the NHS should therefore be spent on optimising data curation. They are, in broad terms, correct. This is a historically neglected space that must be addressed systematically through open innovation and open competitive funding if the nation is to unlock the huge power in NHS data.

The challenge of curation in NHS patient records data

Routinely collected NHS electronic health record data is unlike much bespoke research data, because it was not created explicitly for the purpose of research or analysis. NHS data is typically created for a specific administrative purpose: GP records are largely a "memory aid" for clinicians and patients to help inform decisions about care and, to an extent, guide payment; SUS/HES data is to monitor, or pay for, hospital activity.

Reflecting their origin, individual data points in healthcare often have a much more ambiguous and contextual meaning than operational and logistics data in other sectors. A unit of currency is always consistent. A box of product with a bar code, and its warehouse location, are similarly unambiguous. But a diagnostic code denoting "pre-diabetes" on a patient's record could have a wide range of meanings, in different settings; these codes may be used differently (or not at all) by different clinicians, at different times, in different organisations; and features such as "pre-diabetes" must often be inferred from other traces on a patient's record, such as blood test results, treatments, referrals, or test requests.

In addition, NHS data contains far more granular detail than is needed for a specific analysis. A team wishing to understand the number of children with asthma in each GP practice, and compare the frequency of patients' asthma reviews, does not need to use every detail about every single diagnostic event, measurement, treatment event, or referral event in their final analysis. But they may need access to some or all of this detailed data to create their "analysis ready" dataset, with single variables to denote more broad brush concepts such as "patient has asthma" or "asthma review has taken place".

Current norms around curation: dispersed and duplicative

This curation work can be done well or badly. The historic norm is for it to be completed in an ad hoc fashion, often bespoke for each single analysis, with different technical implementations, methods and tools used by each individual or team; no consistent culture of "Reproducible Analytical Pipelines"; almost no formal culture of sharing; and no real "commons of knowledge" around data curation. This is no criticism of the individuals and teams delivering the work, as it reflects the current landscape of tools, incentives, and collaboration frameworks. There has been almost no open competitive funding for methodological innovation or code on these tasks, limiting the development of better working practices. Previous attempts to bring a systematic approach have largely focused on the low-lying fruit of cataloguing raw data, rather than the substantive challenges around data management; or focused on creating a small number of "assured" variables, usually for some specific managerial task, that address only a small number of use-cases and miss the complexity and diversity in data curation.

A systematic approach to curation built on shared, open code and methods

This challenge can readily be addressed with a systematic approach. Firstly, the system must adopt modern, open, collaborative approaches to computational data science, based on RAP, sharing code (alongside adequate technical documentation) for all data management work. As above, this will help reduce duplication, build a commons of knowledge, and build capacity through reciprocal learning.

Secondly, a small number of teams with a strong track record of open delivery should be resourced to produce curation code on key clinical topics and areas, accompanied by appropriate technical documentation. This should be an open funding call for teams from all sectors to deliver deep dives of curation, validation, and sense checking for 1-3 single clinical topics, in projects co-led by practitioners and developers, delivering open code.

Thirdly, the system should create an Open Library where all NHS data curation work can be shared; and an obligation for all NHS data curation work to be shared here. This should have a dedicated staff with appropriate skills in data science, curation, and technical documentation. It should permit any NHS data user to store information such as code, validity tests, and technical documentation. Code should be shared on a "user beware" basis. It is crucial that any data curation library is not solely a repository of accredited or approved curation code, or the outputs of a small number of pre-selected groups: it should admit all code, but have the facility to display to users which variables have been "assured" by specific organisations, as a tagged subset of all code within the library; and signpost any objective data validation that has been done. Full detailed suggested technical features are given in the full text.

Fourthly, there should be open competitive funding to drive methodological innovation and open code in this complex technical space, in close collaboration with Research Software Engineers, rather than closed approaches to resourcing. Examples of work that UKRI/NIHR could fund includes: work describing the quality and completeness of coding in common NHS EHR datasets on key clinical areas; methods and code for NHS EHR data validation and description at scale; descriptive work on variation in clinicians' coding behaviour between settings; developing and evaluating interventions to improve the quality of coding, focused on specific clinical or geographical areas; optimal methods, tools, and training for codelist creation and related curation tasks; methods for portable representations of complex clinical and demographic phenotypes.

Lastly, all curation work should ideally be conducted in standard TRE settings as this will inherently create portable and re-usable code, tools, and methods.

The destination

Overall this approach will deliver well-curated NHS data for all NHS, academic and life sciences users. It will minimise duplication, harness deep existing expertise across the system, free up analyst time for more innovative work, and improve the quality of curation by surfacing all work for reciprocal review and improvement. A process of "curate as you go, share as you go" will also help to avoid mis-steps of the past, whereby some projects have set out on unrealistic projects to curate all possible NHS raw data - and all possible derivates of it - without prioritising by task, necessity, or practicality. The ultimate goal is that any new NHS analyst, academic researcher, or innovator in the life sciences sector can approach NHS data centres and find a practical, curated library of analysis-ready variables, all adequately documented, and all ready to use off-the-shelf, or review and augment.

None of this should be taken to mean that there should be anarchy. There is an extremely important role for a small number of official standard definitions for purposes such as official national monitoring of specific activities: but this narrow range of official definitions is not the only curation acts that can ever be committed (or usefully shared) in NHS data, whether for service analytics, research, or life sciences. It is vital that any library, code, tools and strategy for NHS data curation admits of the existence of more than only a small number of official "standard" definitions, for a narrow range of variables, created for a narrow range of official users and purposes.

Conclusions

Various projects around NHS data curation have been previously and recently proposed, some with extremely high proposed budgets. While substantial progress can be made with less, the system is correct to have prioritised and valued this work highly. It is wrong to say that NHS data is "dirty", as some have done: it was created for practical purposes in direct care; those using NHS records for an additional new purpose must bear the challenge of reshaping them into something that meets their needs.

Good data curation with open methods is a job in itself; and the key to capitalising on the vast raw data resources that the NHS has collected over the course of 73 years. It will deliver the skills and knowledge to drive the related challenge of interoperability between clinical systems. And it is the bedrock of all subsequent work with data, positioning the UK as a global destination for health data science, delivering the life sciences vision, and using data to improve the quality, safety, and efficiency of care.

Recommendations

The following high level recommendations will help achieve this goal. They map onto detailed recommendations, and background, in the full review text. To inform strategic decision-making in a space that has seen limited progress, the full text also contains a detailed description of what raw NHS records look like, and how these are processed into an "analysis ready" dataset.

12. Stop doing data curation differently, to variable and unseen standards, duplicatively in every team, data centre, and project: recognise NHS data curation as a complex, standalone, high status technical challenge of its own.

    Set up an NHS Data curation planning and delivery team Cur 2.

13. Meet this challenge with systematic curation work, devoted teams, shared working practices, shared code, shared tools, and shared documentation; driven by open competitive funding to develop new shared curation methods and tools, and to manually curate data for individual datasets and fields.

    Detailed recommendations on the shared working practices, shared code, tools and documentation are found in Cur 1, 4, 13, 14 and 16; use RAP principles for curation Cur 1; share all publicly funded data curation code Cur 4; standard tools to convert raw data into analysis-ready datasets Cur 13; portable representations of data management code Cur 14; NHS Digital and others to accept dataset requests in code Cur 16; role of academia in supporting data curation Cur 15, 17-19; open competitive funding call for foundational work on data curation Cur 15; build capacity in clinical informatics through medical curricula Cur 17, universities Cur 18, Cur 19; resource pioneer teams to adopt open curation methods and curate data for all at scale Cur 5; ensure national programmes lead by example Cur 6; resource teams to curate data and share code, methods, validity checks and variables in an open library for commonly used national datasets Cur 7; Run an open competitive funding call for foundational work on data curation Cur 15.

14. Use TREs as an opportunity to impose standards on how commonly used datasets are stored, and curated into analysis-ready tables.

    Use consistent environments to facilitate re-usable curation code Cur 9; require use of national TREs for tasks using national datasets Cur 10; create and enforce consistent standards for local implementations of national datasets Cur 11; curation standards for local TREs Cur 12.

15. Create an open online library for NHS data curation code, validity tests, and technical documentation with dedicated staff who have appropriate skills in data science, curation, and technical documentation; so that new analysts, academics and innovators can arrive to find platforms with well curated data and accessible technical documentation.

    Produce and maintain an open public library of data curation code Cur 3.

Modernising NHS Service Analytics

Good data analysis is at the heart of NHS work to improve the quality, safety, and efficiency of services. Data can be used to compare service activity and clinical outcomes between organisations; to identify opportunities for improving the quality, safety, and cost effectiveness of services; to locate excellence, and share best practice; to model and forecast waiting lists; to predict the best locations and sizes for new services; to evaluate service recovery after the COVID-19 pandemic; to measure the impact of new interventions or new service delivery models; and to ensure value from clinical contracts. These kinds of analyses deliver direct improvements in patient care by identifying problems early, and improving services for all.

Raw data must be managed, curated, processed, analysed, presented, and interpreted before it can generate action.

As is clear throughout this review, data alone does not produce these insights on its own. Raw data must be managed, curated, processed, analysed, presented, and interpreted before it can generate action. This requires a wide range of features to be in place across the system: individuals with strong analytic skills; good training and oversight; data that is accessible; modern data analysis tools; and data that is high quality wherever possible, with any shortcomings documented informatively and accessibly. It also requires senior managers with the skills to recognise good analytics, understand the reports they receive, and pose informed answerable questions to their analytic staff.

The NHS analyst community

Currently the large NHS analyst community contains a wide range of highly skilled individuals, and numerous outstanding and impressive pockets of world-class excellence. However this workforce has become dispersed and isolated over the preceding decades, and now lacks a supportive professionalised

structure. Other government analyst professions each have a head of profession, clear career paths, well-curated continuing professional development training, and various other features of a strong, structured, organised technical profession. The NHS analysts service has almost none of this: no large formal professional body; no clear career pathway with technical job descriptions and associated skills and qualifications; and very little formal structure around initial training or continuing professional development. There is almost no "commons of knowledge"; only small scale conferences run by enthusiasts; barely a single textbook, other than generic data analysis guides from adjacent fields; and no library of methods, workbooks, and code. Where analysts can access training to develop their skills, they feel this is often informal and voluntary, not clearly rewarded; and that career progress only comes from taking on general management roles rather than becoming a more skilled senior analyst.

As a consequence of these structural challenges there is very substantial variation in analytic approaches taken between different settings. There are many examples of excellent work, using modern and open approaches to computational data science, often driven by a single individual or group. But without structures for sharing knowledge this work cannot easily spread. There is a culture of duplicative working behind closed doors, for national and local analytic teams; and a strong reliance on outdated and inefficient means of data management and analysis, using "point and click" tools such as Excel which, though useful for some tasks in an appropriate context, can obstruct reproducibility, transferability, efficient updates, scaling, real-time analytics, and error-checking in analyses, especially when they become the default norm. Lastly there are challenges around the technical setting in which work is done. Analysts commonly struggle to access NHS data, even when there have been substantial investments in local pooled data projects, and they are often prevented from using modern data science tools such as Python or R by local IT constraints.

Building on talent by building a modern profession

There is a pervasive sense of a profession with great potential that is waiting to be unleashed. This change can be rapidly achieved by creating a robust modern career structure around NHS service analytics, modelled on the Government Statistical Service, with clear technical job descriptions at a range of levels. This should include the creation of an Open College for NHS Analysts that coordinates training through openly accessible online resources and in-person teaching, with courses tailored to job descriptions.

Training should emphasise modern open approaches to computational data science, moving from duplicative manual work to writing analytic code and sharing it alongside adequate technical documentation as described above. There is a role for "point and click" tools, and staff who use only those tools (who may have excellent other skills, such as data communication); but using them should be a strategic choice, not a default product of inertia and outdated skills. Due consideration must be given to the broad range of tasks and skills in the NHS analyst profession: from those doing technical data preparation and analysis (who should use RAP); through to those who specialise in tasks such as data communication (who should work alongside those using RAP).

To ensure the spread of good practice the NHS should create an Open Library of NHS Analytics where analysts can share code, documentation and methods that others can review, re-use, modify, and iteratively improve. Analysts should be provided with access to the data, platforms and tools they need, ideally through Trusted Research Environments (TREs) as discussed below. To make change practical, and provide leadership by example, the system should identify three Data Pioneer teams in Integrated Care Systems that can move rapidly to a full TRE and RAP working style. To ensure the best use of data in the NHS, senior leaders from outside the analytic community should be offered training in how to work effectively with analytic teams.

Lastly, the NHS should embrace help from other sectors such as academia and commercial analysts; but collaborate effectively by ensuring that all external work is conducted using modern open working methods, with adequate technical documentation, as per minimum RAP working practices. This should be embedded in boilerplate contract terms, alongside development of new "best practice guidance" for outsourcing analytic work.

Conclusion

The difference between service analytics and academic research is sometimes overstated, alongside suggestions that the working methods, skills and environments should be regarded somewhat or entirely different. It is important to be clear where there are commonalities, and differences. NHS analysts are meeting the needs of customers around practical questions such describing current service activity, or predicting it. Both groups work on similar NHS patient data. Both groups need NHS data to be adequately documented and curated. Both groups might make trade-offs between speed and accuracy, for different projects at different times. Both groups sometimes use statistical modelling. Both groups require an ability to contextualise and communicate information with rushed stakeholders. NHS analysts might sometimes tend more towards simpler descriptive analytic methods; and the full palette of skills required across the workforce might tend more towards data communication or interpretation for non-technical users; but there is no clear reason to regard them as needing entirely different working practices or platforms when working with NHS data. More collaborative work, and collaboration in platforms, will be to the benefit of all.

Recommendations

The following high level recommendations will help achieve this goal. They map onto detailed recommendations, and background, in the full review text.

16. Create an NHS Analyst Service modelled on the Government Economic Service and Statistical Service, with: a head of profession; clear job descriptions tied to technical skills; progression opportunities to become a senior analyst rather than a manager; and realistic salaries where expensive specific skills are needed.

    Detailed recommendations for an NHS Analyst Service modelled on GES and GSS can be found in NHSA 1; job roles NHSA 2, 3; supporting an NHS Analyst community NHSA 4, 5.

17. Embrace modern, open working methods for NHS data analysis by committing to Reproducible Analytical Pipelines (RAP) as the core working practice that must be supported by all platforms and teams; make this a core focus of NHS analyst training.

    Detailed recommendations on finding and amplifying current good practice can be found in NHSA 6, 7; data analysis environments NHSA 22; ensuring NHS IT policies do not obstruct moden working NHSA 23; rationalising national audits, RightCare, GIRFT, and Model Health System NHSA 24; making change practical NHSA 6, 7, 25.

18. Create an Open College for NHS Analysts: this should devise (and coordinate delivery of) a curriculum for initial training and "continuing professional development", tied to job descriptions; all training content should be shared openly online to all; and cover a range of skills and roles from deep data science to data communication.

    Detailed recommendations on training can be found in NHSA 10-14; RAP training NHSA 15, 16; technical team to house and develop continuing professional development resources NHSA 17; training open by default NHSA 18; review curricula NHSA 21.

19. Recognise the value of knowledge management: create and maintain a curated national open library of NHS analyst code and methods, with adequate technical documentation, for common and rare analytic tasks, to help spread knowledge and examples of best practice across the community; use this in training.

    Create and maintain a curated national open library of NHS Analyst Code NHSA 19.

20. Seek expert help from academia and industry, but ensure all code and technical documentation is openly to available to all, procuring newly created "intellectual property" on a "buy out" basis. Commission "Best Practice Guidance" on outsourcing data analytics to cover: where external collaborations can be most helpful; the role of skilled analysts in guiding procurement; common red flags for delivery; and why RAP builds capacity, quality, and continuity of service.

    Detailed recommendations on creating best practice guidance for outsourced analytics can be found at NHSA 26, 27; NHS and academic collaborations on RAP data science for NHS service improvement NHSA 28; audits of organisations and analyst teams NHSA 8; Analytical Capability Index NHSA 9.

21. Train senior non-analysts and leaders in how to be good customers of data teams.

    Create training specifically for senior leaders to help them become better customers for data analysis NHSA 20.

Information Governance, Ethics and Participation

Current delays and frustrations

The research and analytical community is extremely frustrated with the current arrangements around data access. Researchers and NHS service analysts can spend months or years trying to get multiple permissions from multiple parties including: information governance decision-makers in a range of organisations; individual data controllers (including individual GP practices and hospitals); ethics committees in a range of organisations; and more. It is common for large and small analytic projects to be abandoned, as the resource is either spent or lost during the long slow journey to data access. Because of these barriers, important data analyses that could substantially improve the quality, safety and cost effectiveness of care are not being done.

Understanding the barriers

The solutions to this problem are a mixture of the simple and the complex. Researchers, analysts and policymakers all recognise the need for strict regulation to protect patients' privacy and prevent unethical research. EHR data contains the most personal and sensitive information about individuals: access and use should always be carefully controlled. There is room to improve the design of the regulatory system, in particular around duplication of effort: for example, there should be a de-duplication of application forms; and applicants should be present at decision-making meetings to address factual misunderstandings. However, this alone will not address the fundamental challenges; nor will a simple liberalisation of the rules, not least because there is substantial flexibility in the rules already, which are then interpreted cautiously by a range of actors in a range of roles across a range of organisations.

This culture of caution is driven by a range of factors. Firstly, there is an incorrect belief that patients are against data-sharing. Secondly, there is a lack of clarity in the rules, leaving individual decision-makers feeling exposed by the privacy and ethical consequences of each individual access choice they make. Lastly, the needless current reliance by the NHS on less secure methods for data access (principally, disseminating large volumes of pseudonymised but re-identifiable data to multiple destinations) means that each decision to grant access requires a very deep trust in all aspects of every individual analyst or organisation involved.

Using secure platforms and transparency to earn public trust

These concerns can all be addressed by building and using TREs, where there are technical barriers to misuse; where all uses are monitored to ensure all activity remains with the permissions granted; and where all uses are automatically disclosed to earn public trust though transparency and accountability. Detailed evaluations in recent robust Citizens' Juries sponsored by the National Data Guardian and NHSx show that the public understand the concepts behind robust TREs, and strongly support such work.

TREs should therefore be adopted, as discussed in earlier sections: but their use should also be incentivised by developing a two-track approvals process, with far quicker access to data in a TRE, reflecting the reality that data privacy concerns are largely eradicated by this working practice. Decision-makers across the system will feel more confident about granting access when they are reassured that access is being granted through secure platforms rather than relying excessively on deep trust in each individual successful applicant.

Overdue discussions on monopolies, commercial use, performance management, and controllership

In addition to security, four areas of concern were identified that have slowed data sharing and been left largely unaddressed due to a lack of robust, open discussion with the public and/or professionals. The first is the problem of some individuals, teams or organisations wanting to maintain a monopoly over access to data, to meet their own competitive needs: this is largely an unspoken barrier, and commonly hidden behind claims that IG or technical issues prevent data sharing. This must be robustly addressed with an open professional discussion that leads to resourcing and recognition which rewards those who collect data and then share it with a wide range of other users.

The second is concern from some professionals that the NHS records of their patients will be used to "performance manage" them, sometimes in unhelpful or uninformative ways. This must be addressed by robust professional discussion about the benefits of good, positive audit and feedback for quality improvement; and governance that ensures those wasting NHS staff time with misleading performance metrics are themselves monitored, with their access restricted where necessary.

The third challenge is the multiplicity of data controllers in the system: researchers often have to ask for permissions from 6,500 separate GP practices, and 160 NHS Trusts, to access a small number of records from each. This is inefficient, as each sharing choice requires detailed consideration, and it is likely that the degree of oversight in each organisation will vary widely: indeed there are grounds to think that some are excessively permissive; some excessively restrictive; and some inconsistent. This approach would be better replaced by a system whereby organisations can sign up to shared principles and a collective decision-making body that handles all access requests to their data.

The fourth challenge is widespread concern about the ethics of commercial entities having access to NHS patients' data. This is partly driven by the historic use of data dissemination, which means that the ethics of commercial access are mixed up with the separate issue of privacy risks to patients. This can be addressed by using TREs. Notably, TRE working also provides assurance and transparency around the quality and reproducibility of commercial analyses, and all analyses. However the barriers to sharing are also driven by misunderstandings about the important role of commercial innovators. This can only be addressed by a frank, systematic and open discussion with the public, explaining the work that is done with commercial partners, and building a consensus in good faith. Related to this, exclusive arrangements between NHS organisations and the commercial sector should be avoided; and the NHS should negotiate equity in innovations where NHS data is pivotal to development.

Patient and public involvement and engagement

Patient and public involvement and engagement is clearly central to productive and ethical use of data. The most useful, successful, and impactful health data research projects are often those that: design projects with, and for, patients and the public from the outset; involve a diverse range of representatives in every decision, from data definitions, to interpretation and dissemination; listen to (and act on) the advice, feedback, and input of patient representatives; and treat their values, beliefs and experiences as crucial to success alongside well-curated data, performant software, well executed code, or a carefully designed statistical model. Much great work has been done by this sector: modest suggestions are made below and in the full text around ensuring PPIE is done systematically and robustly at a national level on large recurring questions around data usage, alongside the very many smaller projects done in local settings.

Recommendations

The following high level recommendations will help achieve these goals. They map onto detailed recommendations, and background, in the full review text.

22. Rationalise approvals: create one map of all approval processes; require all relevant organisations to amend it until all agree it is accurate; de-duplicate work by creating a single common application form (or standard components) for all ethics, information governance, and other access permissions; coordinate shared meetings when approval requires multiple organisations; have researchers available to address misunderstandings of their project; build institutions to help users who are blocked; recognise and address the risk of data controllers asserting access monopolies to obstruct competitors; publish data on delays annually; ensure high quality PPIE is done.

    Detailed recommendations on rationalising approvals can be found in IG 1 - 6 and 19; create a single form for all varieties of approval IG 1; streamline meetings IG 2; get researchers in the room IG 3; arbitrator for disagreements over access requests IG 4; single map of all approval processes IG 5; unambiguous guidance when approval is not required IG 19; rationalise the rules on posthumous data IG 6; detailed recommendations on how to help NHS analysts, academic researchers, and innovators navigate approvals are in IG 7-8, and 18; two modest Centres for Regulatory Science IG 7; a clinic to help users who are blocked on access IG 8; boiler-plate templates for patient consent IG 18; detailed recommendations on how to ensure PPIE is high quality, informative, and proportionate are in IG 26-30; reflecting sensitivity and scale of projects IG 26; practical guidance and examples of best-practice IG 27; amplifying excellence in PPIE IG 28; consider centrally commissioning PPIE on common causes of concern IG 29.

23. Have a frank public conversation about commercial use of NHS data for innovation, but only after privacy issues have been addressed through adoption of TREs; ensure the NHS gets appropriate financial return where marketable innovations are driven by NHS data, which has been collected at great cost over many decades; avoid exclusive commercial arrangements.

    Detailed recommendations are in IG 23, 24, 25

24. Develop clear rules around the use of NHS patient records in performance management of NHS organisations, aiming to: ensure reasonable use in improving services; avoid distracting NHS organisations with unhelpful performance measures.

    Detailed recommendations are in IG 21, 22.

25. Address the problem of 160 Trusts and 6,500 GPs all acting as separate data controllers: either through one national organisation acting as Data Controller for a copy of all NHS patients' records in a TRE; or an "approvals pool" where Trusts and GPs can nominate a single entity to review and approve requests on their behalf

    Detailed recommendations are in IG 20.

Approaches and strategy: sequencing, scale, and incremental work

The system as a whole has huge potential. NHS data is unparalleled in its breath, depth and power. The academic research community is world class. There are many pockets of excellence throughout all aspects of the system - some buried, some in plain sight - waiting to be amplified. While there are many concrete examples of bad practice - alluded to in this review thematically, and in proposed solutions - all teams and individuals have clearly set out in good faith to deliver.

There are also deep rooted challenges. Medicine both benefits and suffers from being an early adopter of data, as this has created numerous legacy projects: not old software, but old working methods and teams, deeply entrenched, with institutions and networks to perpetuate them. Both the NHS and academia are huge dispersed ecosystems where each constituent organism has its own different requirements, skillsets, priorities, competitive urges and dispositions: this can drive monopolies, and obstruct common solutions. The current narrow incentives around immediate delivery in academia and NHS service analytics make "platforms for all to use" a secondary concern for most people and organisations. As a consequence, money for platforms - the most crucial ingredient needed in the ecosystem today - is often diverted, de-prioritised, or assigned by organisational politics rather than merit. Lastly, and crucially, there is a shortage of technical skills at the coalface, and at the top of organisations where it is needed to guide strategy and detailed action on complex technical issues.

At its worst, the system often seems to hope it can wish these problems away: to procure a single "black box" service that will meet all our platform needs, or analytic requirements, somewhere else, behind closed doors. In reality there is no single contract that can pass over responsibility to some external machine. Building great platforms must be regarded as a core activity in its own right. We must build teams, tools, methods, working practices and code to meet complex technical challenges around health data platforms and curation, as we do with all other complex technical challenges across the whole of medicine.

We have all of the aptitudes, raw data and ambition to excel at this task on a global stage. Achieving success will require a stepwise strategic approach, with small steps in parallel to current workarounds, to prove out new working methods, and build real technical capacity over three years of delivery. After this, we will be ready to re-evaluate our preparedness for a big bang. Repeating the mistakes of the past will help nothing. Building the future will reap a prize of historic proportions across all of service improvement, research, and the life sciences. It requires only that we own the task.

Recommendations

26. Use people with technical skills to manage complex technical problems: create very senior strategic leadership roles for developers, data architects and data scientists; offer leadership training to those in existing technical roles. (Also: train senior leaders in the basics of data analysis, software development, and clinical informatics; but recognise the limitations of that approach).

27. Build impatiently, but incrementally, accepting that new ways of working are overdue, but cannot replace old methods overnight: we must build skills, and prove the value of modern approaches to data in parallel to maintaining old services and teams.

28. Identify a range of "data pioneer" groups from each key sector: three ICS analyst teams; three national quality improvement registry or audit teams; three academic birth cohort or electronic health record analysis teams; and 1-3 national NHS analytic teams. These should be selected competitively as those with the best current technical skills. Resource them to adopt modern working practices (Reproducible Analytic Pipeline working methods in a Trusted Research Environment alongside Research Software Engineer support) and to develop shared re-usable methods, code, technical documentation and tools; this can be in parallel to "business as usual" in their organisation, but should incrementally subsume it.

    Detailed recommendations for practical work supporting "Data Pioneers" to deliver rapid change and capacity are in TRE 37, 45, 52; Data Pioneer academic research teams adopting RAP and TRE working TRE 37; Data Pioneers for RAP and TRE working in research cohorts TRE 39, 45; Pioneers for RAP in data curation Cur 5; Data Pioneer fellowships in NHS service analytics NHSA 6; Data Pioneer analytics teams in ICS and Trusts NHSA 7; Data Pioneer groups for Research Software Engineering Open 28; national programmes lead by example Cur 6.

29. Build TRE capacity by taking a hands-on approach to the components of work common to all TREs. Avoid commissioning multiple closed, black box data projects from which little can be learned, or framing these as "experiments". Experimentation is only powerful where it delivers openly shared working methods, code, outputs and technical documentation from which all can learn.

    - Develop a common "service wrapper" for TRE access, with civil servants.

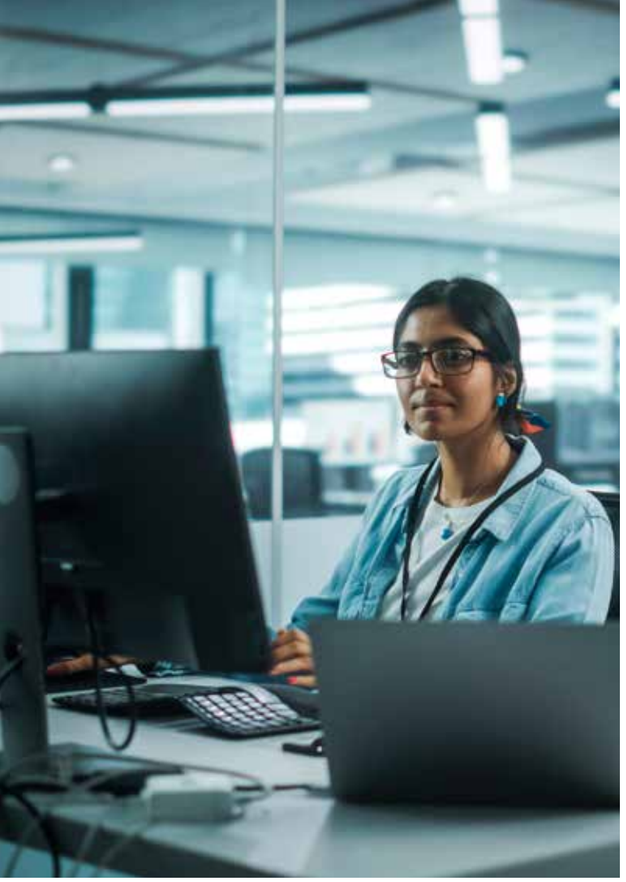
        TRE governance team TRE 11; single standard Service Wrapper model TRE 12; local TRE service model TRE 26.

30. Develop common working practices for the "generic compute and database layer" of TREs with generic skilled technical teams from private and public sectors.

        Detailed recommendations on TRE development are above and in the full text; TRE 54 is especially relevant.

    - Develop "code and methods for working with health data in a TRE" through open competitive funding on key challenges such as data curation, secure analytics, automated disclosure checks, and data minimisation, recognising this as a creative academic and technical challenge requiring deep knowledge of medicine, health data, data science, and software development; ensure all funded work is focused on insights, methods and code that are transferable between TREs and settings.

        Detailed recommendations on TRE development are above. Specific examples of the importance of focusing



on components of the task, rather than procuring a closed “black box” service from academics or another sector, include: create a national standard approach to “output checking” and support automation TRE 13; manage diverse local datasets by creating and sharing standard data curation tools and methods TRE 29; produce and maintain an open public library of data curation code Cur 3; develop standard tools to convert raw data into analysis-ready datasets Cur 13; develop portable representations of data management code Cur 14; run an open competitive funding call for foundational work on data curation Cur 15; open funding calls for projects and programmes around code for health data Open 29, 35, 37, TRE 55.

- Ensure funding for TRE work is competitive, open to all, and overseen by those with data architecture skills; not closed, or prioritised for single organisations who may not have the best ideas and teams.

  Detailed recommendations that include the importance of open competitive funding to amplify talent are throughout, specific examples include TRE 47, 49, 51, 55, Cur 15, Open 29, 35, 37, 38, 41.

- Ensure all TRE teams work in the open, sharing and documenting all code and working methods as they go, to support adaptive innovation.

  Detailed recommendations on open working are throughout. Specific recommendations on TREs themselves being built using open and RAP principles are in Open 45.

- All academic or commercial funding for TREs and code should be openly disclosed including, for each investment: the source of funding; the amount; the recipient; the headline objectives; and a link to the github repository or website where outputs and work in progress can be seen (including code, technical documentation, or live services).

  TRE 47.

31. Focus on platforms by resourcing teams, services and institutions who are focused solely on facilitating great analytic work by other people, working closely with users. Data curation, secure analytics, TREs, libraries, RAP training, and platforms are the key missing link: they will only be delivered if they become high status, independent activities.

    Detailed recommendations on putting platforms first are throughout the text and recommendations.

Conclusions

The high level recommendations in this summary document give an overview of the key risks and opportunities. The full text of the review contains detailed background and practical recommendations, reflecting the deep technical complexity of this space.

The NHS has a phenomenal resource in the detailed data that has been collected for tens of millions of patients, over the course of many decades. This data represents a spectacular opportunity to improve NHS care, and drive innovation in the life sciences sector. It is also a research resource of global importance, not least because the NHS population is larger - and more ethnically diverse - than other countries with similarly detailed health records.

We should all regard it as a profound ethical duty to make the best use of this resource. 73 years of NHS patient records contain all the noise from millions of lives. Perfect, subtle signals can be coaxed from this data, and those signals go far beyond mere academic curiosity: they represent deeply buried treasure, that can help prevent suffering and death, around the planet, on a biblical scale.

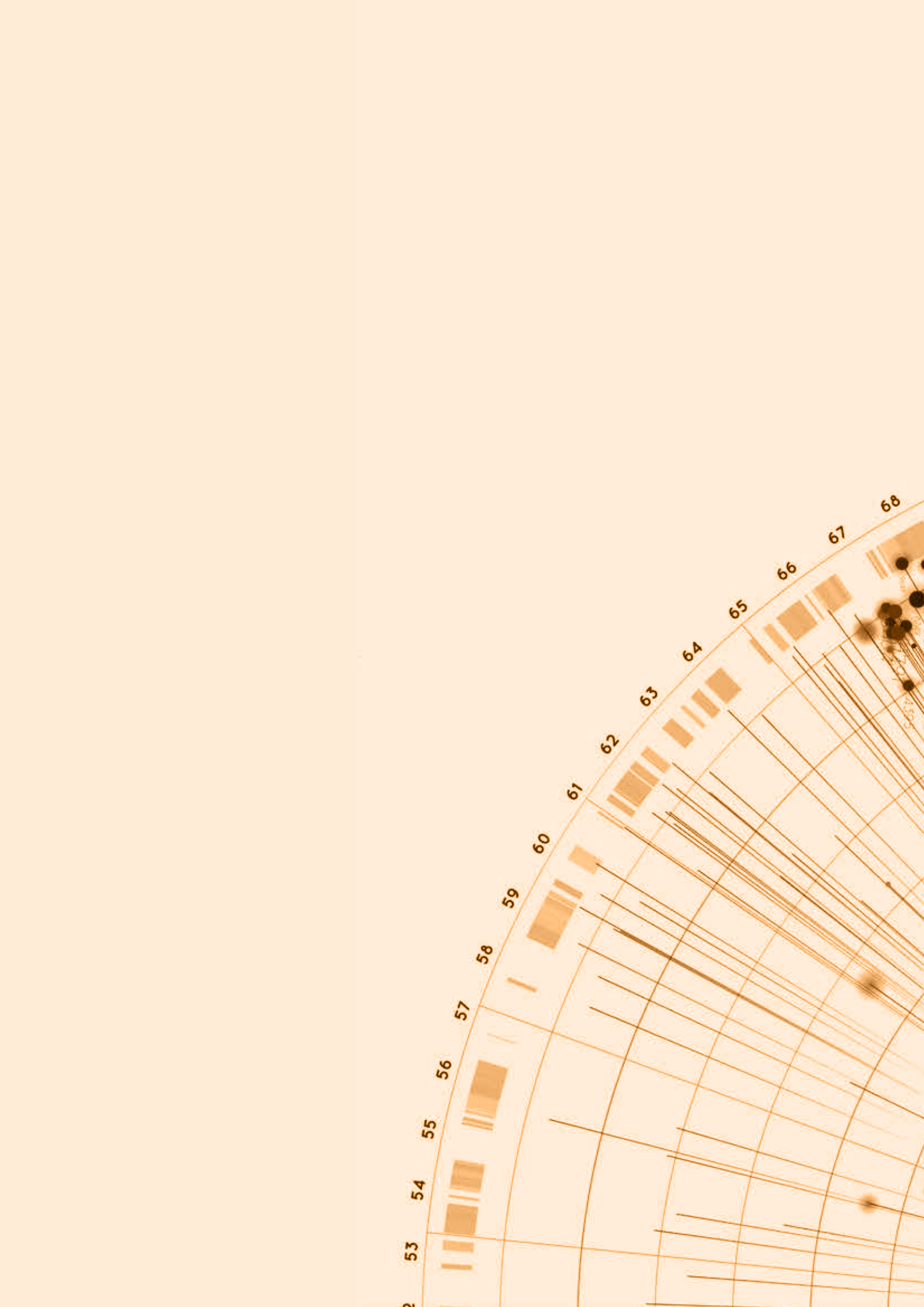
In the past, there has been a tacit tendency to view NHS data almost as a free lunch: as if the cost of sharing 60 million health records was little different to putting some files on a USB stick. In reality, modest strategic investment is needed to ensure that this complex data is well curated, and shared in platforms that are both secure, and performant. This can be done efficiently, but only by accepting the technical complexity of the work; adopting modern, open working practices; and using open, competitive funding to create a thriving technical community that drives better use of data through only shared methods and code. Building capacity and platforms may take three years; but it has been put off, unhelpfully, for much longer. To continue with current working practices means accepting a huge hidden cost of duplication, outdated working methods, data access monopolies, needless risk and, above all, missed opportunities.

By investing in a coherent approach to data curation, and a small number of secure platforms, the nation can unlock all the untapped potential in NHS data. Any investment in this space will pay phenomenal dividends. For less than the cost of digitising one hospital the system can have the secure data platforms and workforce needed to realise the full value of NHS data.

This will reap rewards across the global research community, where NHS data is an unparalleled resource, and where we already excel at delivering smaller, single academic research projects. It will drive innovation across the whole life sciences sector, where our data, platforms, and workforce could lead the world. And it will drive change across the NHS, where smart use of data can help improve the quality, safety and cost effectiveness of all care, for all patients.

In all this, we must earn public trust. NHS data is only powerful because of the profound contribution of detailed health information from every citizen in the country, going back many decades. If we can show the public that we have built secure platforms for data sharing, then every patient can confidently embrace sharing their records, safely and securely, for the good of the NHS, and humanity, around the globe.

COVID-19 has brought fresh urgency, and shone a harsh light on some current shortcomings. But future pandemics and waves may bring bigger challenges; and there were always lives waiting to be saved through better, broader, faster, safer use of NHS data.


68
67
66
65
64
63
62
61
60
59
58
57
56
55
54
53