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

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

April 2022

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.

Contents

Review Team

Professor Ben Goldacre, Goldacre Review Chair

Professor Ben Goldacre, Goldacre Review Chair Director, Bennett Institute for Applied Data Science; Professorial Fellow, Jesus College; Bennett Professor of Evidence-Based Medicine, Nuffield Dept of Primary Care Health Sciences, University of Oxford

Ben Goldacre is a clinical researcher at the University of Oxford where he is Director of the Bennett Institute for Applied Data Science, and Bennett Professor of Evidence-Based Medicine in the Nuffield Department of Primary Care Health Sciences. He advises government on better uses of data and leads an academic team that uses large health datasets to deliver research papers and tools including OpenSAFELY.org (a new model of secure analytics platform that runs across unprecedented volumes of linked NHS patient data); OpenPrescribing.net (an open data explorer for NHS GP prescribing choices with over 20,000 users a month); and TrialsTracker.net (an open tool that monitors clinical trial reporting performance). He is also active in public engagement: his books including 'Bad Science' have sold over 700,000 copies in more than 30 countries; and his online lectures have over 5 million views.

Jessica Morley, Goldacre Review Researcher

Policy Lead, Bennett Institute for Applied Data Science, Nuffield Department of Primary Care Health Sciences, University of Oxford; Wellcome funded DPhil Candidate, Oxford Internet Institute, University of Oxford

Jess is a social science researcher at the University of Oxford where she is the policy lead for the Bennett Institute for Applied Data Science, and a Wellcome funded DPhil candidate at the Oxford Internet Institute. Her health data policy work is supported by the Mohn-Westlake Foundation. Prior to moving into academia full-time, she was a civil servant for the Department of Health and Social Care and latterly, NHS.

Nicola Hamilton, Goldacre Review Secretariat

Civil Servant, Department of Health and Social Care

Nicola is a Civil Servant in the UK Health Security Agency, an executive agency sponsored by the Department of Health and Social Care. She has worked in the Civil Service for the last six years, across a range of projects and programmes, most recently focusing on health data.

Senior Stakeholder Group

Over the course of the review, we have been guided by the below listed colleagues, comprising our Senior Stakeholder Group (SSG). The purpose of the SSG throughout the review was to provide guidance to the review team, ensure that the right questions were being asked, the right people were being involved, and that the content met the expectations of the Secretary of State and the needs of the system. In short, the members of the SSG have acted as critical friends throughout. We have greatly appreciated their time, encouragement, and constructive criticism throughout. Their involvement in this capacity does not imply that they are responsible for, or in agreement with, the full content of this text.

Matthew Gould - CEO, NHSX

Simon Madden - Director of Policy & Strategy, NHSX

Dr Louise Wood - Director of Science, Research and Evidence, Department of Health and Social Care

Jem Rashbass - Executive Director of Data Services, NHS Digital

Jackie Gray - Executive Director of Information Governance, NHS Digital

Dr Felix Greaves - Director of Science, Evidence and Analytics, National Institute for Health and Care Excellence

Professor Sir Ian Diamond - UK National Statistician, ONS

Dr. Arjun Dhillon - Clinical Director & Caldicott Guardian, NHS Digital

Professor Sir John Bell - Regius Professor of Medicine, University of Oxford

Ming Tang - Chief Data and Analytics Officer (interim), NHS England

Dr Colin Wilson - Deputy Director, Innovation and Growth, Office for Life Sciences

Hadley Beeman – Chief Technology Advisory to the Secretary of State, Department of Health and Social Care

Background information

Terms of reference for the review

1. How do we facilitate access to NHS data by researchers, commissioners, and innovators, while preserving patient privacy?

2. What types of technical platforms, trusted research environments, and data flows are the most efficient, and safe, for which common analytic tasks?

3. How do we overcome the technical and cultural barriers to achieving this goal, and how can they be rapidly overcome?

4. Where (with appropriate sensitivity) have current approaches been successful, and where have they struggled?

5. How do we avoid unhelpful monopolies being asserted over data access for analysis?

6. What are the right responsibilities and expectations on open and transparent sharing of data and code for arm's length bodies, clinicians, researchers, research funders, electronic health records and other software vendors, providers of medical services, and innovators? And how do we ensure these are met?

7. How can we best incentivise and resource practically useful data science by the public and private sectors? What roles must the state perform, and which are best delivered through a mixed economy? How can we ensure true delivery is rewarded?

8. How significantly do the issues of data quality, completeness, and harmonisation across the system affect the range of research uses of the data available from health and social care? Given the current quality issues, what research is the UK optimally placed to support now, and what changes would be needed to optimise our position in the next 3 years?

9. If data is made available for secondary research, for example to a company developing new treatments, then how can we prove to patients that privacy is preserved, beyond simple reassurance?

10. How can data curation best be delivered, cost effectively, to meet these researchers' needs? We will ensure alignment with Science Research and Evidence (SRE) research priorities and Office for Life Sciences (OLS) (including the data curation programme bid).

11. What can we take from the successes and best practice in data science, commercial, and open source software development communities?

12. How do we help the NHS to analyse and use data routinely to improve quality, safety and efficiency?

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. You will see them quoted throughout.

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

Chapter 00

Executive Summary

Goldacre Review

Summary

Scope

This review was tasked with finding ways to deliver better, broader, safer use of NHS data for analysis and research: more specifically, it was asked to identify the strategic or technical blockers to such work, and how they can be practically overcome. It was commissioned to inform, and sit alongside, the NHS Data Strategy. The recommendations are derived from extensive engagement with over 300 individuals, 8 focus groups, 100 written submissions, substantial desk research, and detailed discussion with our Senior Stakeholder Group.

The untapped power and potential of NHS data

NHS data represents an exceptional and globally important resource. For 73 years the NHS has collected detailed records and data, on tens of millions of patients, from a huge and ethnically diverse population. Because of this diversity, analytic outputs created from NHS data can help save lives around the world. The combined GP records of the nation, as just one example, cover every person in the country; they go back many decades; and they capture some information for nearly every contact with health services, with huge detail on prescriptions, treatments, blood tests, referrals, and diagnoses.

This dataset - the full medical history of millions - contains almost unfathomable depth and potential. Data is at the core of all good work in healthcare. Data is how researchers learn which treatments work best, and for which patients. Data has driven the global response to the COVID-19 pandemic, and can help target work on the post-pandemic backlog. The life sciences sector can use data to evaluate and refine medications, or develop whole new classes of medical technology. By monitoring all activity and outcomes, NHS analysts can find new opportunities to improve the quality, safety, and cost effectiveness of care, across the whole health service.

The importance of platforms

The nation has world class researchers, and outstanding raw data. But raw data does not do great work on its own. This data must be curated, managed, cleaned, reshaped and prepared by people. Then it must be made available in well-designed platforms, which earn public trust through security and transparency, and which facilitate sharing and re-use of prior work.

At present the system relies on multiple small data projects that do not join up, distributing large volumes of the same patient records to an uncountable range of very different sites for different projects and teams. This duplicates implementation costs, data preparation costs, governance costs, and risks; it fosters monopolies, and obstructs transfer of ideas and analyses between settings. It obliges the system to rely excessively on weak security practices such as “pseudonymisation” (removing names and addresses from detailed health records) without always acknowledging the shortcomings; and to build complex systems of governance, contracts and trust that can only manage the security risks inherent in data dissemination by acting in a slow and risk averse manner. This approach has arisen from decades of “getting by”: but it can never scale to the kind of access needed for a world leader in data science.

Building practically for the future

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. The full text of this review contains detailed background and practical recommendations, reflecting the technical complexity of this space. The high level recommendations below give an overview of the key risks and opportunities. The system should act now, starting with small teams of Pioneers

to capitalise on existing pockets of excellence, building capacity and new ways of working in parallel to old approaches; after this, a full transition can come quickly.

This is a generational opportunity. We need a brief, rapier-like focus on platforms, creating teams and ideally institutions who are tasked solely with facilitating analytic work by other people. 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, driving research, health service improvement, and innovation. COVID-19 has brought fresh urgency: 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.

Summary recommendations

Platforms and security

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 40, 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

7. Promote and resource "Reproducible Analytical Pipelines" (RAP, a set of best practices and training created in GDS and 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 31-34; 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 and Knowledge Management

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.

NHS Data Analysts

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 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.

Governance

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 Patient and Public Involvement and Engagement (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 agreements.

    Detailed recommendations are in IG 23, 24, 25.

24. Develop clear rules around the use of NHS patient records for 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

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.

    - 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; ensure this is focused on insights, methods and code that are transferable between TREs.

      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 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.

30. 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

In the past, “data infrastructure” meant beige boxes in large buildings. In the 21st century, data infrastructure is code, and people with skills. As noted in previous reviews, many shortcomings in the system have been driven by a “destructive impatience”: constantly chasing small, isolated, short-term projects; at the expense of building a coherent system that can deliver faster, better, safer outputs for all users of data.

If we invest in platforms and curation - at less than the cost of digitising one hospital - and engage robustly with the technical challenges, then we can rapidly capitalise on our skills and data. New analysts, academics and innovators will arrive to find accessible platforms, with well curated data and accessible technical documentation. The startup time for each new project will shrink, productivity will rocket, and lives will be saved.

73 years of complete NHS patient records contain all the noise from millions of lifetimes. 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. It is our collective duty to make this work.

Chapter 01

Modernising NHS Service Analytics

Goldacre Review

Summary

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.

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.

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.

Background

The work of NHS analysts

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 which settings have the greatest opportunities to improve care. The same analyses can be used to identify excellence, and help organisations that have achieved outstanding performance to share insights or strategies with their neighbours. It can be used to model and forecast waiting lists, beyond simple counts of queues. It can be used to predict the best locations, and sizes, for new services. It can be used to evaluate whether new interventions or reorganisations have achieved their clinical or logistic objectives, and monitor volume of activity alongside cost to ensure value from clinical contracts. These kinds of analyses deliver direct improvements in patient care by identifying problems early, and improving the efficiency of services for all.

Data alone does not produce these insights: 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: it requires individuals with strong analytic skills; good training and oversight to ensure analysts are using the best methods for the right analyses; senior managers who have the skills to recognise good analytics, understand the reports they receive, and pose informed deliverable questions to their analytic staff; data that is accessible; and data that is high quality wherever possible, with any shortcomings documented informatively and accessibly.

NHS service analytics can sometimes seem deceptively simple, but it often entails technical work well beyond the immediately obvious. For example, organisations will commonly want to understand waiting lists for interventions - well outside of any media discussion on the topic - as part of the routine everyday activity around planning and understanding services. This information is used to change the number of sessions used for a given activity in each setting, and so on.

Data alone does not produce these insights: raw data must be managed, curated, processed, analysed, presented, and interpreted before it can generate action.

Understanding the waiting time for a given activity in a given organisation may seem superficially simple, but important methodological and judgement calls must be made around the best calculation: for example, using "total time waited by people who have been seen in the past month" may not reflect current changes in the waiting list, as it will look only at those who entered the list earlier in time; whereas using a "census" measure, calculating average waiting time so far for people waiting on the list, will over-represent people who have waited a long time, because the people who were seen swiftly disappear from the census sample (and they may be more urgent cases, or systematically different from other patients in other ways). The key question is likely to be "how long will new entrants wait": but this

will commonly entail a degree of “modelling” around future scenarios, because it requires an understanding of the current throughput of the service, and an ability to forecast the impact of changes in future throughput, and future changes in the rate of new referrals. Many of these analytic challenges become even more complex during COVID-19: for example, there has been a substantial downturn in referrals for services, but it is very challenging to know how many of those apparently missing referrals will return over the coming year.

This example speaks to a wider challenge: to capitalise on opportunities to improve health and care, we need the data and outstanding data analysis, not just in academia, but at the clinical coalface of the NHS, generating insights that help clinicians and key decision makers make informed choices that directly improve care for millions of patients across the NHS.

Workforce and Training

Organisational Structures

The team discussed NHS analytics with a wide range of NHS service analysts, from a variety of national and local organisations, at junior and senior levels. It is clear that there have been substantial changes over the preceeding decades in how NHS service analysis is delivered, and how the workforce is organised. Overall our very strong impression was that the NHS analysts community contains a wide range of highly skilled individuals, and numerous outstanding and impressive pockets of world-class excellence, but that the workforce has become very dispersed and isolated over the preceeding decades. As context for this, over the same period of time data science has become one of the most high status, open and visible professions in the global jobs market; and the analytic functions more broadly other parts of government have also developed higher status and a range of strong organisational structures around their work.

It is important to note that this workforce is very dispersed, both geographically and organisationally. NHS analysts are dispersed widely across the many diverse national, regional and local organisations of the NHS. At national organisations there are analysts in NHS England (in numerous teams and directorates, many of which have their own specific analyst teams); in NHS Digital (in various teams and roles); and in numerous smaller organisations such as the NHS Business Services Authority. More distantly, NHS service analytics is also at the heart of work in national organisations such as the Care Quality Commission, using the same underlying data to monitor similarly vital clinical activity and outcomes. Analytic work is also commonly given by various diverse national and local organisations to various diverse external or adjacent organisations including Commissioning Support Units, and some university groups, alongside large and small private providers of analytic services. In local organisations, there are analysts in Integrated Care Systems, individual hospital Trusts, Clinical Commissioning Groups, and to varying degrees in GP Federations, Primary Care Networks, Academic Health Sciences Networks, and other organisations.

This workforce is also very large. There are certainly thousands of staff currently working in analytic roles across the NHS, but the actual number of analysts is unclear, in testament to challenges around visibility and structure of this workforce. The Health Foundation recently estimated that there are approximately 10,000 people working in analytic roles across national and local NHS organisations.

At present there is no formal professional body (but some outstanding grass roots organisations); very little formal structure around teaching and training, or “continuing professional development”; and a general lack of clear technical job descriptions or qualifications specific to NHS service analytics that managers without technical skills can use to evaluate the skills mix that they have, or need, in their service. This stands in very striking contrast to arrangements around the many other technical professions that drive NHS activity including laboratory technicians, clinicians, nurses, physiotherapists, radiographers, and so on. A strong illustration of this challenge can be seen in how NHS analysts are located in the administrative structures of the NHS, and its employment banding: under the standard NHS Terms and Conditions of Service (“Agenda for Change”), NHS analysts are classed as “admin/clerical” staff rather than “scientific/clinical”.

“There's no way for people who work in clinical informatics to prove themselves to be legitimate to the system. It's not professionalised."

- Interviewee

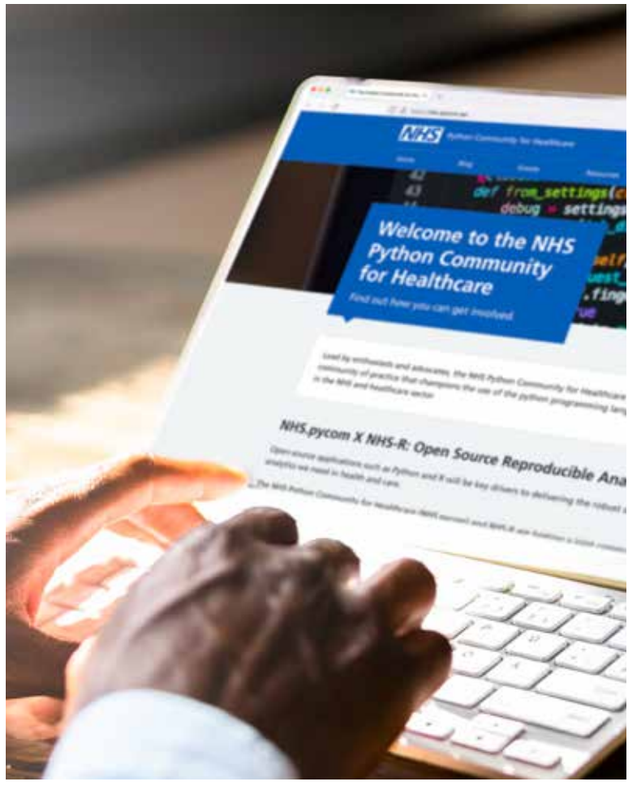
By contrast, other technical specialties in the NHS have a strong and diverse strategic infrastructure to guide their initial training, career pathways, supervision, recognition, ongoing training, and to help create a technical “commons of knowledge” around their work. This will include Royal Colleges or other professional membership organisations that are typically high status and well resourced; detailed job descriptions at a range of seniorities; formal arrangements nationally and locally around training both at entry level and for continuing professional development; and so on. Similarly other government analyst professions such as the Government Economic Service, the Government Statistical Service, and the Government Operational Research Service each have a head of profession, with clear career paths and progression opportunities, supported by well-curated continuing professional development, and the other features of a strong technical profession. These national organisations set out clear best practice guidance, offer analysts accreditation, and require analysts to adhere to a code of conduct. These models for technical work in both the NHS and government provide a clear template for future work around the NHS analysts service.

The Government Analysis Function

The “analysis function” in wider government has faced similar challenges to the NHS analysis community. The government professions within the Analysis Function include: Digital Data and Technology Profession, Government Actury’s Department, Government Economic Service, Government Geography Profession, Government Operational Research Service, Government Social Research, Government Statistician Group. Collectively these groups represent approximately 17,000 people involved in the generation and dissemination of analysis across government, including statisticians, data scientists, researchers, economists, policy experts, data journalists, data visualisation experts, methodologists, and more.

The network was created to facilitate the sharing of best practice, provide consultancy services, build capability, create tools, guidance and standards, and monitor performance of the different analysis functions. The Government Analysis Function Career Framework, for example, was collaboratively developed by all the analytical professions and is designed to describe typical analytical roles across government, including the main skills required to perform each role at varied skill level. Where specific skills are highlighted, the framework signposts to relevant training available such as that available for: data visualisation; communicating insight; quality assuring analysis; data management; data modelling data cleansing, and data enrichment techniques; statistical methods; and software programming, tools and techniques.

By contrast, the NHS analysts community has an array of small-scale grass roots organisations, typically run on modest subscriptions for a small number of participants, or with small amounts of intermittent charitable funding from organisations such as the Health Foundation. Strong examples of this include the NHS-R community, the NHS-python community, and the Association of Professional Health Analysts. These organisations do an excellent job of championing the work of NHS analysts, and providing them with opportunities to exchange ideas, but they need dedicated support to scale, and cannot begin to match the strategic and structural role of the organisations serving other similar technical and analytic professions.



The NHS-R Community

R is a widely used statistical and graphics programming language. The NHS-R community was set up in 2018 as a joint project between the University of Bradford, the Yorkshire and Humber Academic Health Science Network Improvement Academy, the Association of Professional Healthcare Analysts and NHS Improvement and NHS Wales, and funded by the Health Foundation’s Advanced Analytics programme. In keeping with the aim of the Advanced Analytics programme to improve analytical capability across the health and care system, those running the NHS-R community (led by Mohammed A Mohammed, Professor, University of Bradford & Principal Consultant at the Strategy Unit) aim to promote and enable the use of R in the NHS to improve data analysis and develop shared solutions to common analytic challenges.

Today, the NHS-R community - which is hosted by the Strategy Unit and now also has support from the NHSX Analytics Unit - runs the premier data science conference in the NHS, along with regular skill-based webinars; has an active blog where members share ‘R-tips’; has over 1000 members in a thriving slack community offering help and support to each other; runs the NHS-R Academy which offers free training courses to the NHS, recognises the contribution of its members via honorary titles (eg Fellow, Champion, etc) and develops and shares R-based solutions which address common problems whilst sharing all its resources on GitHub; produces a podcast where community members and guests discuss the use of R and open analytics more broadly.

The NHS-R community - with its welcoming and supportive ethos - provides an excellent entry point for those looking

to begin learning the benefits of working with modern, open and collaborative data science tools, and those looking to further develop their skills by accessing the expertise in the community. R is also a good place to start learning how to conduct analysis using a script-based language as it has excellent documentation, numerous libraries, and a worldwide user base that freely shares learning and resources.

The more recent appearance of the NHS python community is testament to the appeal of the approach to the broader NHS digital data and technology, analyst, and IT workforce. However, its continued growth and development relies on the time and commitment of volunteers who contribute to the community in addition to ‘doing the day job.’ This, understandably, limits the scope - in terms of what it covers (for example, data analysis more than data management) - and scale - in terms of the number of analysts the NHS-R community are able to reach and connect with.

Association of professional healthcare analysts (AphA)

AphA is another largely grass-root-led organisation that aims to raise the profile of healthcare analysts by providing them with: support, professional development opportunities, networking opportunities and learning opportunities at conferences, access to resources which highlight best practice, regional branch events and webinars where members can share knowledge. AphA also offers members the chance to become professionally registered with the Federation of Informatics Professionals (FEDIP) as a means of validating their analytical skills and competencies.

More recently, AphA has been collaborating with NHSx to scope the need for an NHS analyst competency framework (the discovery document can be accessed here) as part of NHSx's work programme on developing the analytical profession. This is part of a wider programme of work being overseen by the 'Developing Data and Analysis as a Profession Board' and commitments outlined in the Data Strategy, and supported by AphA, to achieve their joint goal of helping to develop analytical capacity and capability in health and care fit for the 21st century, ensuring equitable recognition of analytics as a profession with progressive career pathways.

The fact that these organisations have run for so long, on modest resource, with comparatively large impact, demonstrates that there is a strong need in the NHS analytics profession for strategic structures, training, and so on. However they also show that this cannot be delivered solely through "inspiration" to the profession on its own. The individuals involved in driving these organisations have achieved phenomenal outputs but they do not have the scale, voice, access and infrastructure of the substantive structures in other professions. They should be closely involved in all subsequent work around the NHS analyst service, the impressive power of smaller scale voluntary and charitable work has now been clearly demonstrated, but the limit of the voluntary model in this space has also been met.

A range of proposals on better structures for the analyst workforce are given at the end of this chapter.

Training, salaries, and a "Commons of Knowledge"

Technical professions typically have a range of deep technical descriptions for the expected skills at each level of the workforce, each tied to opportunities for continuing professional development and career progression. The NHS service analysts community were clear that there is very little for their work which can compare to the range and depth seen in adjacent NHS technical specialties, or government analytic functions.

Training is largely patchwork, and frequently framed as an informal or voluntary activity, with no clear certification of skills, or recognition of skills in formal career pathways. Broadening out from the experience of individual analysts, it is clear that there is almost no "commons of knowledge" around NHS analytics. For most technical specialties and professions one would expect to see a diverse range of textbooks, journals, libraries of analytic code, CPD-accredited courses, conferences devoted to training and detailed technical illustrations of new approaches, and so on. For NHS analytical work there is almost none of this: barely a single textbook, other than mostly generic data analysis guides from adjacent fields; very few formal courses; no substantive work around accreditation for training; small scale conferences run by enthusiasts; no library of methods, workbooks, and code; and so on.

Alongside this, NHS analysts repeatedly told us that they can only progress by being seen to take on management roles, rather than by becoming more senior, productive, and technically capable. This partly reflects the categorisation of analysts by NHS employment structures as "administrative/clerical", as management roles are more appropriate as a sole criteria for progression in those roles. However in all other technical and analytic professions, while managerial duties may be one route to seniority, it is not generally the sole route.

"Nobody wants to work in an NHS trust on NHS data, it's a nightmare and we can't pay people appropriately."

- Interviewee

For context, data scientists and data analysts with qualifications or a proven track record are currently some of the most in-demand employees in the global jobs market. In the private sector, they can expect to earn in excess of £80,000 per annum, with higher salaries as they develop more productive skills. By contrast, advertised NHS analyst salaries are typically between £25,000 and £45,000, often in job descriptions requiring deep knowledge in industry-standard open source data science tools such as R and Python, which command extremely high salaries in other sectors. While many individuals will take a reasonable pay-cut for public service (especially with the geographic inelasticity of NHS salaries across the country) the scale of this salary disparity is unrealistic, especially when staff face the prospect of working in a hidden and poorly recognised corner of the NHS.

AnalystX

AnalystX is a data and analytics community of practice, created at the start of the pandemic to support collaboration throughout the health and care analytical community. The community now has 16,000 members, committed to a common vision of data driven, evidence-based decision making by sharing learning across health and social care beyond typical organisational and geographic boundaries

AnalystX is a community run by volunteers, for the community, with committed support from a wide range of health and industry strategic partners, working together to build the data and analytics community through 5 approaches:

1. Educate through community-curated and indexed data and analytical resources such as dashboards, web applications, evidence syntheses, insight reports, best practice case studies and guidance
2. Support through hosting weekly analytical case study and learning and development webinars, and evolving sub communities forming virtual cross-organisational teams which come together to solve common challenges
3. Cultivate by assisting analysts to start and sustain their learning through creation of the ALX, with 53 free on demand learning modules covering technical, data, software and people skills, and the promotion of free training opportunities offered by strategic partners
4. Encourage by promoting the work of members through discussion forums enabling people to contribute directly to the site, and setting up a champions network drawn from the community to contribute to and promote the community
5. Integrate by encouraging members to use their new knowledge for real change in their own work, tying together recognition of learning and site contributions through digital badge awards

This new initiative is very welcome: a range of detailed proposals around augmenting work are given at the end of this chapter.

Analytic approaches and Reproducible Analytical Pipelines

As a consequence of the structural and organisational challenges outlined above, it is clear that there is very substantial variation in analytic approaches taken between different settings. There are many outstanding examples of excellent work, using modern and open approaches computational data science, often driven by a single individual or small group in one setting. But these pockets were largely invisible to those outside of their group or organisation. It is clear that there is also a strong reliance across the system on more outdated and inefficient means of data management and analysis, using "point and click" tools such as Excel which undoubtedly have a role but can commonly obstruct reproducibility, transferability, efficient updates, scaling, real-time analytics, and error-checking in analyses.

Health Foundation

The Health Foundation is an independent charity committed to bringing about better health and health care for people in the UK. They have a long history of using, and championing the use of, data analysis to help them achieve this aim. This includes, more latterly, conducting analysis in house and sharing the code openly on GitHub; providing funding to improve the analytical capability across the health and care system via the 'Advancing Applied Analytics' grant programme, the NHS-R Community and the Association of Professional Healthcare Analysts; sharing examples of analytics excellence on Twitter every Friday; and encouraging collaboration between analytical teams across the UK via the recently established Networked Data Lab.

None of this should be taken as any criticism of any individuals: rather they are a direct consequence of the broader challenges in the organisational structures around individual NHS service analysts. Indeed there is a clear picture of excellent work being done, but then neither captured nor spread: while there are many examples of ad hoc sharing between individuals, there is a lack of a structured strategic approach to analysts developing and then sharing best practice or innovation around specific analytic tasks such as data extraction, data management, data curation, data analysis, data presentation, or specific analytic challenges for NHS service analytics within those over-arching themes.

In the chapter on Open Working there is a detailed description of Reproducible Analytical Pipelines. This is a powerful brand first developed in 2017 by the Government Digital Service to describe a range of contemporary best practices for data management and analysis in the public sector. It is built around a single core principle: "At any point in the future we should be able to look back at this work and be able to reproduce everything that we have done today - something that is difficult with manual and semi-manual processes." RAP emphasises a range of working principles. It promotes the use of open source languages such as R and Python rather than proprietary tools: this ensures that all subsequent users are guaranteed to have access to the same tools; and reflects the emphasis placed by the open source software community on good documentation, flexibility, and extensibility, which are all powerful principles for all data analysis. RAP is now a very strong, very broad movement across government departments with extensive training and deep experience of implementing change in diverse settings.

The NHS can and should rapidly adopt RAP working practices, both for service analysis and for research. More detailed proposals on this are given below; the importance of modern, open approaches to analytics are considered in the chapter on Open Working.

Public Health Scotland: RAP in a health context

Public Health Scotland represents a powerful example of how an organisation can rapidly modernise its approaches to data management, and rapidly embrace an approach built in the principles of RAP. Over the course of 2 years they swiftly modernised their data centres, and staff, moving from slow manual approaches to RAP. This harnessed a range of opportunities: it reduced the scale of duplication; and made updates and amendments to data-driven reports substantially faster and more efficient. It improved morale, and helped analysts focus more on high value work. It also played a valuable role in capacity building: where new analysts can find, read, re-use, and modify the previous notebooks made by others in their team containing code and data (such as Jupyter, or R-markdown) then they can learn swiftly with real, applied examples. This was all made more possible by the fact that most PHS work can now be done in a common data environment, rather than a range of implementations of smaller datasets, or cuts of data, on individuals' machines.

Analytic environments

NHS analysts repeatedly expressed frustration around the technical platforms in which they are required to work. A more detailed description of the challenges and solutions in this space are given in subsequent chapters on Open Working, Data Curation, and Trusted Research Environments. A brief overview of the specific challenges faced by NHS analysts is given here in context.

Analysts are commonly faced with a range of diverse computational environments which are either slightly different, or very different, to those in other national or local organisations. This means that their data management or analysis code is not readily portable, and that work cannot be readily re-implemented in adjacent settings. Related to this, the same national datasets such as GP data, or Secondary Uses Services (SUS) and Hospital Episode Statistics (HES), are often stored in different local and national data centres in slightly different ways: sometimes these are small differences, such as column headings; sometimes there are modest differences, such as different versions of languages such as SQL needed to interrogate the raw data; sometimes there are huge differences in the data model, and the extent or manner of its pre-processing prior to its arrival, such that work can barely be engineered to be transferable.

NHS analysts often found they struggled to access the modern computational tools they need, such as Python or R, as their smaller local IT team did not have the skills to implement these, or felt they were outside of what they could securely approve: the team was told as a consequence of this, by more than one analyst, that in their experience it is common for people to use workarounds outside of the approved working practices of their workplace, on the basis that they had evaluated an alternative option themselves and felt it to be reasonable, in the light of what they felt to be unreasonable or uninformed IT policy choices that would block their work (see also TREs). There is also a very widespread frustration at the enormous amount of time spent on data curation (see Data Curation) and the lack of sharing and transferability of this work between settings because of the divergences in data structures, skills, and computational environments.

Lastly NHS analysts in national and local settings described their very deep frustration at not being able to access data in a timely manner, for a range of different reasons. These barriers included 'Information Governance (IG) problems, where the rules prevented them being able to access data. More specifically, analysts expressed a sense that decisions around IG were often slow, but also arbitrary, and could sometimes reflect "conflicts of interest" whereby other organisations inside and outside the NHS would selectively grant, or not grant, access according to their own strategic or competitive preferences. Specifically it was very surprising to find local or regional analysts describe how they were unable to access large and widely promoted data aggregation projects that would help them meet their analytic needs, saying "I've checked with my boss, I've just been told, that's not for us". These challenges around access and analytic environment represent a clear barrier to the delivery of efficient analytic services, that can be swiftly improved. A range of proposals on better structures are given in the sections on Data Curation, Open Working, and TREs.

NHS Managers

Analysts do not operate in a vacuum. They are typically tasked with answering analytic questions for operational purposes such as improving the quality, safety and efficiency of local services. Some generalist managers and clinicians, with little access to analytical training, feel out of their depth when commissioning or evaluating analytic insights provided to them. Conversely, some analysts express frustration at senior leaders asking unrealistic questions, or wanting to view numeric outputs as concrete 'indicators' rather than practical 'measures' to initiate a discussion. Both spoke of the need for better dialogue for collaborative development of good outputs.

"We don't have enough leaders with informatics backgrounds. Analysts are some of the smartest people I know, they need to understand how the NHS works as well as the data analysis experts and those skills aren't permeating upwards."

- Interviewee

Some technical knowledge - alongside clear job descriptions and certification - is also necessary for managers to be able to understand the difference between different types of analysts, recruit well, and guide strategic issues such as team size, composition, and focus. While there is no need for senior leaders to consistently have deep technical knowledge, it is clear that skills in this area are very variable between organisations. Analysts need to be managed by staff with appropriate data literacy, to feel confident in those managing them, to know that they are being asked to answer important and influential questions, and to know that they have the backing and support they need to be bold in their analyses. Strong, supportive, and informed leadership can also help with staff retention. A range of proposals are given at the end of this chapter on offering appropriate analytics skills to senior leaders.

"New ways of working aren't championed at a leadership level to use analysts. Part of the reason for high turnover of staff who are specialist data scientists, etc. We can get these people in but can't retain them."

- Interviewee

Clinicians and national audits

Concerns were expressed by clinicians that they were unable to access analytic services, and in some organisations and areas there seemed to be a sense that analytic work is principally for senior managers to monitor financial activity. While this is not optimal, it may reflect local culture, or individual access. However, overall the structures of data usage for clinical service improvement are fractured, with a range of overlapping local, national, and single-topic projects. These include RightCare, Model Health System, Getting It Right First Time (GIRFT), and a very diverse range of single topic audit projects built around "registries", which are typically bespoke data collections and extracted databases focused on a narrow band of clinical activities.

As with regional and national NHS service analysis the individuals working on these projects have a wealth of detailed, impressive, and vital domain knowledge around the clinical meaning of the data, its strengths and weaknesses, the best means to manage and interpret it, and so on. Many of them have invested an enormous part of their professional working lives into creating and improving these services. At present, however, for each of these projects separately there is a tendency for all aspects of data collection, extraction, management and analysis to be done inside one single organisation or team, which is run as a "full service" arrangement, delivering a finished static output such as an annual report at the end.

This structural approach, with a series of parallel "city states" is largely an accident of technical history, reflecting an era when this was the only natural or practical working style for such datasets. The individuals in these projects are typically highly skilled and collaborative, and many work hard to involve others in their own work. However, in the modern context this structure risks being duplicative, does not always produce the best "commons of knowledge", and risks creating or reinforcing monopolies around access to data and knowledge that can block innovation and high quality analytics for patient care. This stands in contrast to a more dynamic and open approach where data is collected, stored, and given detailed technical documentation for all to see and understand, with access granted to multiple competing and collaborating teams, who all then work in the data to generate analytic insights. A range of proposals in the TRE chapter around delivering a more open, competitive and collaborative ecosystem built around TREs with shared code and documentation; with due consideration of the need to preserve current skills, knowledge, outputs and incentives.

External Collaborators

At present NHS organisations commonly outsource analytic work from commercial providers, or "NHS adjacent" organisations such as Commissioning Support Units. This can be a very efficient way for a local or national service to access skills that it needs only intermittently,

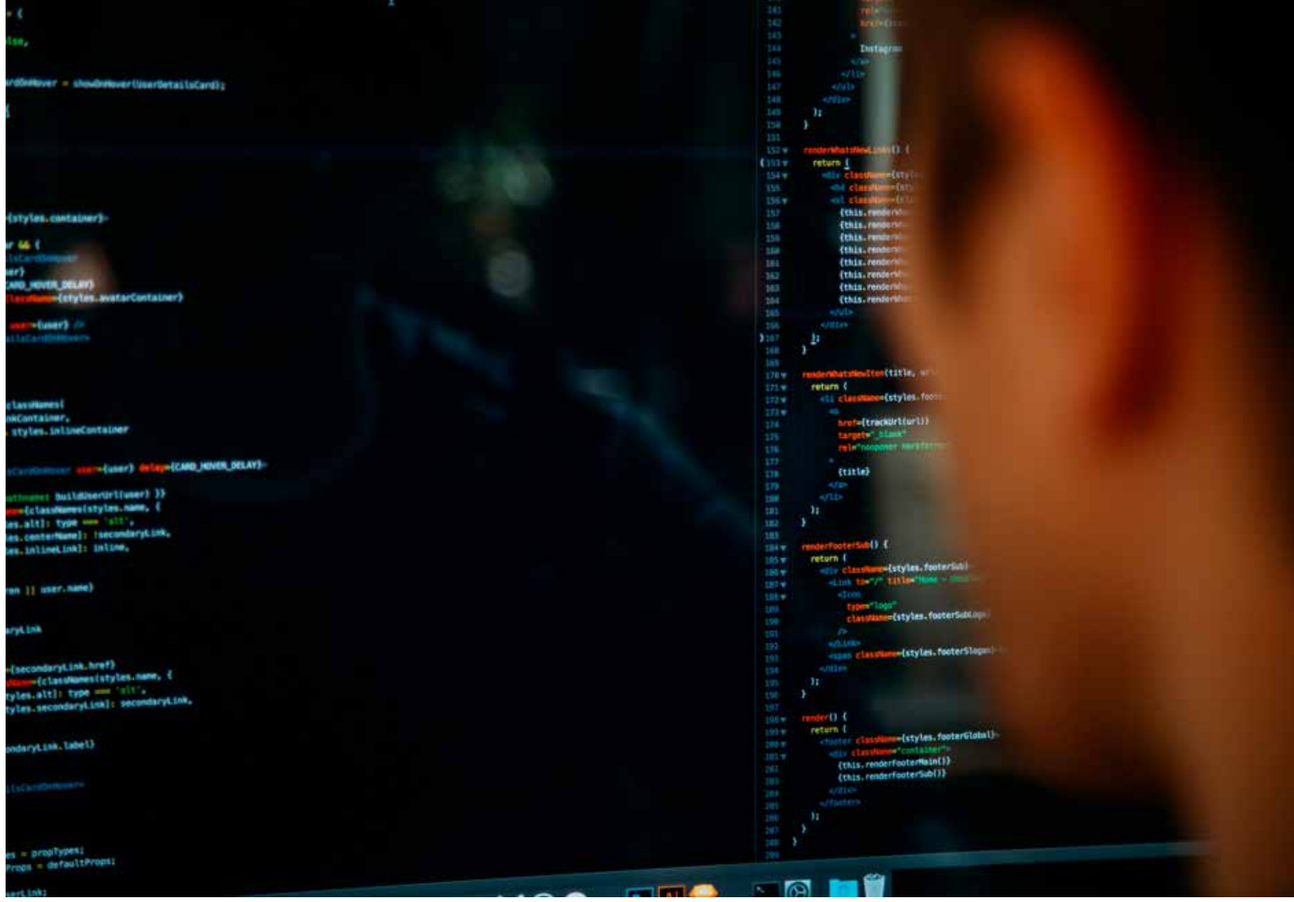
or cannot deliver itself. However, a number of analysts had very critical views on this practice. They felt that outsourcing of analytics happened in a less strategic manner, reflecting a lack of skills and knowledge in their own leadership team, managers' vulnerability to procuring services on the basis of an appealing PowerPoint slide deck, or lack of knowledge about which aspects of work are challenging. For example, various analysts shared situations where an external provider had promised to deliver a range of graphs, dashboards and reports, but had then returned to the same service's own NHS analysts instructing them to provide all the relevant data extracts, pre-processed, reshaped, explained and interpreted. In the analysts' view this meant that NHS analysts were doing the very work that the outsourced provider had been paid to do on the basis of a lack of skills in the NHS analysts' team.

It is not possible to adjudicate on stories such as these, but it is useful to think through the strategic causes, advantages, and disadvantages of using commercial collaborators to deliver analytic work. Regarding causes, the current low status, low visibility, and low pay of the NHS analytic workforce may well lead to situations where the local NHS in-house offer is less appealing than a strong presentation from a commercial firm with experience in winning contracts. This can be addressed by building the analytic workforce and by placing an emphasis on data communication and presentation skills in NHS analysts' training, job descriptions, and CPD. Similarly, where there are situations in which NHS leaders are outsourcing work without understanding the nature of the activity, this can be addressed through better training in analytic skills for NHS managers, as discussed.

The benefits of working with external commercial partners are clear. They may have true economies of scale from which a local organisation can benefit. They may have strong domain knowledge from adjacent fields of business analytics. They may be able to employ highly skilled data scientists, data analysts, and software developers on realistic market salaries, where NHS organisations themselves are more limited, paying market rates for accountants, clinicians and legal expertise, but struggling with administrative barriers to pay realistic salaries for specialist skills that have emerged over more recent decades.

It is important to also have a clear view of the risks. There is more general concern about reliance of the public sector on outsourced contractors, which is a generic strategic and political question that lies firmly outside the interest of this Review. More specifically for NHS data analysis, this is detailed technical work requiring generalist data science skills and deep domain knowledge around the clinical context, alongside the strengths and weaknesses of NHS EHR and administrative data. Where contractors lack that knowledge, they may not give good service; where they have it, or develop it, there is a risk of it being captured, at a time when there is increasing recognition of the problems caused by closed recent approaches to data analysis in the NHS, and the need for a rich, technical Commons of Knowledge in this space to drive innovation by all.

Related to this, the team encountered situations where individual outsourced contractors were asked to conduct data curation tasks where the information was not shared, and therefore could not be evaluated for accuracy, quality or safety. Similarly, the team encountered descriptions of large databases containing large volumes of NHS patients' detailed electronic health records data being held by external contractors for their own internal use in providing analytics as a "full service" to the NHS. In the sections on Open Working, and Trusted Research Environment, there are a range of proposals to avoid knowledge and patient data being captured in siloes by public and private organisations: addressing these challenges will be crucial to drive an open, innovative, productive ecosystem or analytics and research. At the end of this chapter are various proposals on guidance for managers on best practice around outsourcing NHS service analytics.



Recommendations

The NHS analyst workforce is a crucial part of the health service, with vast potential waiting to be tapped in numerous energetic pockets of excellence across the country. Below are a range of practical proposals around training, working methods, and organisational structures for the profession. These relate closely to recommendations in subsequent chapters: the need for open working methods; the need for secure and efficient analytic environments; and the need for more efficient proportionate processes around information governance.

Professional Structures

There is a clear need to adopt the structured approach in other technical and analytic professions.

NHSA 1. Create an NHS Analyst Service modelled on GES, GSS, GORS

The system must capitalise on the dispersed talent throughout the NHS and let inspiring individuals lead their colleagues. The Government Economic Service, Government Statistical Service, Government Operational Research Service and Government Social Research Service provide an appropriate model for a new body. These professions each have a head of profession, clear career paths and progression opportunities supported by continuing professional development. They hold their staff to high standards by setting out clear best practice guidance, offer accreditation, and set out a clear code of conduct. This service should be responsible for delivering most or all of the following tasks, set out in recommendations NHSA 2 - NHSA 9.

NHSA 2. Create clear job descriptions for NHS analysts at a range of levels

In collaboration with NHS analysts, Association of Professional Health Analysts, NHS R Community, Royal Statistical Society and the Cabinet Office Central Digital and Data Office, this proposed NHS Analyst Service should create clear job descriptions for NHS analysts from entry level to head of profession. These should be used nationally to clarify roles, recruit staff, and help senior managers (who likely lack technical skills) to identify, appoint, and train data analysts. The job descriptions should be underpinned by a clear competency framework outlining the specific technical skills required to complete specific technical tasks. Such tasks may include: data communication, data analysis, data management, statistical modelling, risk prediction or service evaluation. The government functional standard provides examples of high-level descriptions including analyst, analytical assurer, analysis commissioner, senior officer accountable for analysis in an organisation. The Digital, Data and Technology Profession Capability Framework provides a good foundation to build on, with detailed and specific examples for different levels of data analyst, data engineer, data scientist and performance analyst. Analysts need clear career paths to seniority that allow them to become senior highly skilled analysts rather than generalist managers.

NHSA 3. Revise Agenda for Change, and ensure technical staff are paid realistic salaries

NHS analysts, software developers, engineers, and other technical staff should no longer be classified as "administrative / clerical" staff. Technical roles require their own category within the Agenda for Change pay scale framework, with their own job titles, capabilities, competencies, KPIs (Key Performance Indicators), and competitive remuneration packages. The NHS must stop expecting to pay highly skilled technical staff in data science and software development on salary scales devised for low and intermediate level IT technical support. Data scientists and software developers in the commercial sector routinely earn more than their manager, customer, or commissioner: this reflects market value, and is no different to employment of senior clinicians, accountants, lawyers, or other technical specialists by organisations. If barriers are hit when discussing offering higher salaries to senior developers with longstanding experience, the anchor point for negotiations should be NHS clinicians' salary.

NHSA 4. Support an NHS Analyst Community

Learning from existing community building and CPD activities, including that conducted by medical school deaneries and NHSx, and in collaboration with key organisations such as APHA, NHS-R community, Royal Statistical Society, ensure NHS analysts have access to a range of community building activities, including regional and / or organisational CPD groups, such as the RAP meetups run by the Government Statistical Service.

NHSA 5. Develop an annual data conference for NHS service analysts

This should be a high-status event with training, presentations, awards, possibly as part of NHS Expo, giving NHS analysts an opportunity to come together, learn, share, and celebrate examples of excellence, create a community, and raise the status of data analysis across the health and care system. The conference should be held during work hours and should be free to attend.

NHSA 6. Find good staff, and empower them quickly with "Data Pioneer" fellowships

The system has a challenge: to rapidly foster the development of complex new behaviours and teams. It is hard to meet this challenge through central edicts; during the review it has become clear that there are many pioneers throughout the system who are already exhibiting the desired working practices, often without recognition or support. In other parts of medicine, the system uses "fellowships" to give independence, status and influence to clinicians with specific desired skills in analysis, teaching, or research. This is a powerful opportunity to find people with strong existing analytic skills, using RAP or open methods, and rapidly empower them in practical terms. This should be an open competitive programme where applicants can seek resource to cover half of their salary for 3 to 5 years so that they can spend half of their time spreading and developing their working methods, teaching, developing teaching materials, or receiving analysts for supervision and mentorship on placements.

NHSA 7. Identify three "Data Pioneer" analytics teams in ICSs and Trusts

As per current policy, the future structure of regional analytics and service management will revolve around Integrated Care Systems (ICSs), each covering a population of approximately 1-4 million patients, and analysts in NHS Trusts. To demonstrate the power of modern open methods in NHS service analytics, the NHS Transformation Directorate should identify three Integrated Care Systems and/or hospital trusts where there are strong existing skills in analytics, informatics, and/or software engineering to act as Data Pioneer teams. 2-4 individuals from each group should be provided with advanced training in modern, open, computational and collaborative working methods, including RAP, with the rest of the team given training in the foundations so that they can learn from 'doing' under the direction of the group leaders with advanced training. These Data Pioneer teams can lead by example, providing open documentation of their work for others to learn from, make the methods and code local service analytics more visible to the wider community, and feed into the wider programme of modernisation around the NHS analyst service. It may be useful to choose teams and individuals who are close to working with raw NHS records data, as they will have substantial internal knowledge around data management that will be widely applicable.

NHSA 8. Commission intermittent code and analysis audits of organisations and analyst teams for service improvement

In collaboration with academics, and key organisations such as AphA and the NHS-R community, the proposed NHS Analyst Service or NHS Analyst Head of Profession should commission regular code audits of all organisations that have received public funding for health data research or analysis, including funding for the development of intermediate knowledge objects (such as re-usable code, documentation, or functions). These audits should follow a set methodology; be published openly; and be used for the explicit purpose of improving performance, rather than penalising poor performance. Specific criteria should be developed in collaboration with the community but include: delivery and use of open code; open methods; open data where possible; sharing insights; support for CPD in work time; whether staff meet JDs with training, CPD or other proof. Good performance should be further incentivised, by highlighting best practice examples.

NHSA 9. Create an Analytical Capability Index

This should be developed independently, and used nationally, to track whether individual organisations have room to improve and signal to leadership where gaps lie in their organisation, how they compare to peers, who they can learn from. Careful consideration should be given as to how best to present the results, and whether this should be made public or not. It is important that the results are only used to drive genuine improvements, and not used for arbitrary contextless performance management.

Training

There is a need for a strategic and structured approach to training of NHS service analysts.

NHSA 10. Create an Open College for NHS Service Analysts

This brand will emphasise that the training is open to all interested parties, and that the analytic methods promoted are themselves modern, open approaches to data science. This Open College should contain the following activities, set out in NHSA 11 - NHSA 21.

NHSA 11. Devise the content of a national training programme for NHS analysts: initial and CPD

Clear job descriptions and pathways must be tied to training and, where appropriate and non-onerous, proof of competencies. Health data is complex, as are health services: working as an analyst in this setting requires a range of specific knowledge around practical health data analytics, alongside more general technical skills in data management, analysis, and visualisation. The NHS Analyst Service should be tasked with devising a curriculum and training requirements for the key competencies associated with job roles, with clear recognition of existing experience or training in and outside of health, and so on. This should be facilitative rather than restrictive, and be focused on informing high quality training, rather than imposing onerous requirements to gather paperwork as proof of skills.

NHSA 12. Oversee funding and delivery of training, both open online and one-to-one

Having identified the training required, this must be delivered and resourced. Training pathways must be more than an ad hoc list of standalone links to existing online resources. Training should be an appropriate blend of openly accessible online training, such as MOOCs, accompanied by formal one-to-one or group work to support feedback, supervised practical work, and evaluations, in the situations and skillsets where this more expensive in-person training can be shown to deliver better outcomes than open online work alone. Both MOOCs and in-person training should be resourced through a framework where providers can compete to receive funding and offer training as in other sectors. This should include a range of activities at a range of levels including: core training through new post-graduate certificates, diplomas, and degrees in applied practical analytics for health and social care; and Continuing Professional Development (CPD) opportunities for in-career analysts, consisting of refresher courses, opportunities to learn new skills, and so on. CPD courses should award completion certificates, proof of CPD, recognised or even required by managers, and these should be matched where relevant to competencies in analyst job descriptions. This should be overseen by a governing body and developed in close partnership with AphA, RSS (Royal Statistical Society), and the NHS-R Community.

NHSA 13. Establish new core training for analysts

Replace the Health Education England Graduate Management Training Scheme in health informatics and health analysis specialisms with a specific graduate training scheme in health data analysis that should include: core training; rotation in different parts of the NHS (for example, in primary care vs. secondary care); the opportunity to specialise (for example,, in data engineering vs. data management vs. data analysis); and specific training in the use of modern open computational methods. This will require funding and coordination from national Arms Length Bodies (ALBs), local NHS organisations, national funders, NHS Leadership Academy, HEE, academic organisations (where they can demonstrate a specific commitment to practical NHS service analytics) and more.

NHSA 14. Outline clear, non-onerous CPD training requirements for analysts

Once CPD opportunities have been made available, it should be a requirement that all analysts gather CPD points. Progress should be evaluated annually and tied to progression opportunities to ensure that participating is appropriately incentivised. This process should be as light-touch as possible, for example, confirmation of attendance at a conference or completion of an online module. Expectations regarding the amount of CPD points required, and the type of activity that 'counts' should be adjusted accordingly to job description and level of seniority. This training cannot be delivered by simply linking out to generic data science training resources from other suppliers covering work in other sectors outside of health, albeit that these may well be fertile starting points for modification into bespoke training on RAP and computational methods for NHS data.

NHSA 15. Embrace RAP and modern, open working methods

Excel has its place, and training will be required at a range of levels for a range of skills. However, there is a clear need to move away from inappropriate use of inefficient and outdated "point and click" methods for analysis, and towards a model based on Reproducible Analytical Pipelines with modern, open, collaborative approaches to data science. Intermediate, and advanced analyst training should focus on enabling the workforce to develop skills in modern, open, collaborative computational data science with an emphasis on reproducible analytic pipelines covering concepts and skills such as R, version control, GitHub, Jupyter notebooks, Pandas, and similar. This does not mean that everyone in the system must become an expert software developer: but it does require some changes in skillsets and emphasis. RAP has a proven track record in other parts of government and in Public Health Scotland, with a strong model for spreading change in organisations. These will be new skills for many and so training will entail more than links to external generic data science guides. Training should be practical and include completion of tasks inside sandboxes so that mistakes can be made safely. The training provided by the ONS Data Science Campus provides an excellent example, and training should be developed in close collaboration with the RAP teams. The chapter on Open Working discusses these issues in more detail.

NHSA 16. Ensure training focuses on RAP as much as Machine Learning

There is a tendency for training to be diverted into more exotic forms of data analysis such as Artificial Intelligence or Machine Learning. These have their place, and there are many existing resources that meet these training needs very well outside of health analytics for those who

have already developed outstanding skills in data science. However, the key unmet training need in the service is RAP, and the delivery of analytics using modern, open methods to achieve improvements in efficiency, sharing, quality, transparency, documentation, and reproducibility. This must be the priority for any training programme.

NHSA 17. Create a technical team to house and develop continuing professional development resources

Training in technical skills needs to be delivered by those with technical skills in data science as applied to NHS data. It cannot be delivered by generalist data scientists alone. Training also needs to be kept up to date. The aim should always be to provide training in the most advanced computational data science skills; what these are will change over time. Providing a team of technical specialists with adequate funding to develop, deliver, share, and curate training, including the development of tools such as sandboxes where analysts in training can practice their coding and analysis against dummy data that reflects real NHS data, will be essential if training is to be high-quality and up to date.

NHSA 18. Ensure all training is open by default

The traditional funding model for training from universities and many other providers is to charge per-attender. Wherever online training resources are commissioned they should be open by default, with all video lectures, training materials, written content, exercises, and code shared openly. This may result in a somewhat higher unit cost for teaching resources procured on a buy-out basis but will represent a better investment in the medium term. It is necessary to remove access barriers to knowledge and training, in a space that urgently requires up-skilling, and to avoid imposing a requirement on analytic staff to ask permission of generalist managers, who may lack technical skills themselves, for access to training budgets that require onerous engagement with bureaucracy. This is particularly important in a complex ecosystem such as the NHS where behaviours, expectations, resourcing, management styles and training availability may vary widely between local and national NHS organisations. Fully open access to all NHS analyst training resources will also create substantial network benefits. It will make these training resources accessible to NHS staff in adjacent specialties who wish to up-skill, including managers and clinicians, enabling them to drive forward better use of data in their own teams and organisations; and to outside elements from the public and private sector making it clearer to them how the NHS uses data to improve care, and how they can interact to offer help and support, or improve analytic work with better tools, algorithms, services, or insights.

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

Hire a team of 10 people to create an open library of code and workbooks for key recurring tasks, examples of best practice, ‘how-to guides’, code for common analytical queries, codelists, variables, and so on. It must be unashamedly technical but meet the needs of staff with a range of abilities. The library should be presented as a flexible open online website, with clear tagging, careful thought around discoverability of resources, and careful curation of individual resources into “training arcs.” The library delivery team should be led by an editor experienced in producing good open online technical resources; it should include analysts but also include expertise in technical writing, knowledge management, online education, and publishing. An MVP (Minimum Viable Product) should be created within 6 months by pulling together the best existing resources from national and local teams in close collaboration with all key stakeholders and teams already listed above. This library should be closely tied to (and feeding into) online teaching and CPD. CPD points should be provided for contributing to the library and there should be an obligation for any analyst developing code with public resources to contribute the outputs to the library for re-use. The need for better knowledge management around code and methods for NHS data analysis is also discussed in the sections on Open Methods and Data Curation.

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

There should be an expectation that non-analyst staff, especially those managing analysts, have sufficient data literacy to conduct informed conversations about data. This will require that non-analyst managerial staff and clinicians have access to training to help them make better use of data in their day-to-day jobs; enable them to make smarter decisions about how to use data for performance management; enable them to ask better questions of their analysts; and to provide them with the skills they need to distinguish between high-quality and poor-quality analysis. Undergraduate and postgraduate training for clinicians and managers should include knowledge of how data is captured and analysed to improve care. Funders and employers should resource collaborative teaching and training between analysts and clinicians/managers. This training will bring the management of NHS service analysts into alignment with the Government functional standard for analysis. It should ensure that non-analyst staff are, at a minimum, able to confidently evaluate: whether commissioned analysis is compliant and appropriate for intended use; the risks, limitations, and major assumptions of particular analytical methods; whether the output is appropriate for the analytical need.

NHSA 21. Commission a rapid review of medical school curricula and similar

The content of the curriculum in medical schools, allied health professional training, and clinical post-graduate training is always hotly contested, with a wide range of competing communities advocating for their speciality to receive more prominence. Nonetheless there are good grounds to believe that clinical informatics and data science for service improvement are under-represented. A rapid review of current curriculum content in all medical schools and a range of other clinical training will identify where there is need for augmentation and should aim to recommend how training in essential technical skills can be incorporated without compromising other aspects of the curriculum. This can be done by Health Education England.

Platforms and data access

Analysts need access to data in platforms that support modern open working.

NHSA 22. Improve the provision of data analysis environments

In the sections on Trusted Research Environments and Information Governance there are a range of detailed recommendations to ensure that local service analysts have access to data in environments that support modern, open approaches to computational data science, and to ensure that data is not unreasonably withheld.

NHSA 23. Revise NHS IT policy for analysts to ensure it is fit for purpose

Analysts need to be able to use modern computational data science tools such as python, GitHub and docker on their NHS computers. Current IT policies often block the use of such tools. A similar challenge has been faced and recently overcome by the analytic community in government outside of health. This must be addressed in national and local IT policies with clear statements on assurance and risk from the NHS Transformatin Directorate to local decision makers, to make it the norm for work to be delivered using modern computational data science tools and avoid the apparently prevalent problem of analysts using these tools outside of the formal permissions and policies of their workplace. Many but not all of these challenges will be met by delivering better national and local infrastructure for data access; however, there will likely still be a role for individual local machines that facilitate the use of standard modern tools.

NHSA 24. Rationalise national audits, RightCare, GIRFT, and Model Health System

As described above these projects are presently implemented as “full service” arrangements where all data collection, extraction, management, and analysis is done inside one organisation or team in order to produce an intermittent single output such as an annual report. These teams have deep knowledge around their datasets, and the clinical context. However the current working style risks duplication of effort (and risk) around data extraction, data hosting, and data management; blocks sharing of detailed technical knowledge and code around methods for data analysis and outlier detection; does not use the most efficient methods to produce a commons of knowledge; and risks creating or reinforcing monopolies around access to data and knowledge that can block innovation and high quality analytics for patient care. A better model would be for all these services to operate in common analytic environments; where all have access under reasonable constraints; and where all code is shared alongside documentation. This is best delivered through identification of a small number of Data Pioneers among these national audit projects, who are ready to embrace RAP working methods and work in a national TRE or similar to deliver their outputs. This is discussed in more detail in the chapter on TREs.

NHSA 25. Make change practical

The NHS should identify three Data Pioneer ICSs that can move to a full TRE and RAP working style within 6 months; and three Data Pioneer national quality improvement audits (at least one within NHS England) that can move to full TRE and RAP working within 6 months.

External collaborations

External collaborations with the private and public sector are valuable but should be handled thoughtfully with an emphasis on open delivery.

NHSA 26. Commission and promote best practice on outsourcing analytics

It is reasonable for NHS organisations to sometimes seek help from external commercial and public sector organisations to improve the productivity or scope of their analytic outputs, especially in a period of transition while the NHS analyst profession is being developed. However as discussed above this work brings risks around quality, transparency, and development of wider open knowledge on NHS data, whether the external partners are commercial or academic. The NHS Transformation Directorate should coordinate the development of Best Practice guidance on outsourcing to cover the range of scenarios where such external collaborations are and are not beneficial to the system, boilerplate contractual requirements, and red flags around working methods and delivery.

NHSA 27. Require all outsourced or external work to comply with RAP and open working methods

Currently when analytic projects are outsourced to consultancies, academic collaborators or other agencies it is common for only the results to be reported, for example in a PDF or slide deck, without the accompanying methodology or code used to conduct the analysis. This prevents the NHS from error-checking the work, learning from it, or being able to replicate it internally, whether in the organisation that originally commissioned the work or elsewhere in the system. This creates duplication of work, and the loss of knowledge that can create efficient analyses and drive innovation. This cannot solely be addressed by asking external partners for “training” or more detailed narrative descriptions of the methods used. As discussed in the chapter on Open Methods, all NHS data management and analysis code should be accompanied by adequate technical documentation alongside the code, as required by the minimum standards of RAP, openly available for re-use and external scrutiny. All outsourced work should adhere to this requirement.

NHSA 28. Support NHS/academic collaborations on RAP data science for NHS service improvement

UKRI/NIHR should consider running an open funding call specifically for academic teams to collaborate with national or ICS NHS data analysis teams on using RAP and modern open data science techniques to improve the quality of NHS care, to deliver specific outputs, and to build mutual relationships and capacity building around applied analytics. The targeted outputs should be a range of Jupyter notebooks or similar with well-documented open code describing - with appropriate technical documentation - how NHS data was prepared, analysed, and used to identify or address opportunities to improve NHS clinical activity or outcomes.

Chapter 02

Open Working

Goldacre Review

Summary

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 (ONS) and the Government Digital Service (GDS) 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 in GDS and ONS 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 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 start-up 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 (RSE) 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 nobody 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.

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.

Background

Reproducible Analytical Pipelines

Throughout the interviews with senior and junior stakeholders it became clear that the system is in a period of transition, whether in academic or NHS service users of NHS data. This was particularly clear when talking with data scientists or analysts who had approached health data from other sectors. They repeatedly expressed how surprised they were to find that approaches regarded as standard in other parts of industry or academia - such as sharing code, or the everyday working practices of collaborative software development - were not yet the norm in teams working with health data.

The analytic communities in other areas of academic research and government have already recognised and energetically addressed the need to embrace modern, open approaches data management and analysis. In the chapter on Data Curation there was a brief description of Reproducible Analytical Pipelines, developed and implemented by Government Digital Service, the Office of National Statistics, and the analytic professions across government. Here it is useful to explore its purpose and practical aspects, in order to understand how this approach can best be implemented across NHS service analytics and health data research.

The various texts on RAP describe the prior norms around statistics production in government, in similar terms to current working practices in health seen during this review. "Broadly speaking, data are extracted from a datastore (whether it is a data lake, database, spreadsheet, or flat file), and are manipulated in a proprietary statistical software package, and possibly in proprietary spreadsheet software. Formatted tables are often then 'copy and pasted' into a word processor, before being converted to pdf format, and finally published... This is quite a simplification, as statistical publications are usually produced by several people, so this process is likely to be happening in parallel many times... [Quality assurance is then] a manual process which can take up a significant portion of the overall production... as any changes will require the manual process of production to be repeated."

By contrast to this manual approach, Reproducible Analytical Pipelines deliver the same work more efficiently and reliably, using commonplace contemporary practices that have been developed over time by the analytic, data science and software development community. The adoption of standard working practices from the software development community is important, as it reflects the reality that data analysis is implemented by writing code. RAPs reflect a modern, open, collaborative and software-driven approach to delivering 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 the following criteria (adapted from Government Statistical Service):

- Minimise manual steps, for example copy-paste, point-click or drag-drop operations. Where it is absolutely necessary to include a manual step in the process this must be documented.
- Be built using open source software for data management, analysis and visualisation which is available to anyone, preferably R or python.
- 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 (sharing data itself is a separate issue from sharing code, as discussed below, and should be handled very differently).
- Guarantee an audit trail using version control software (such as Git, or in services such as GitHub) which systematically track exactly who has made which changes or contributions to the code, which characters and lines were modified, when, and - as appropriate - why.
- Follow existing good practice for quality assurance.
- Deepen technical and quality assurance processes with code review by peers.
- Contain well-commented code and have documentation embedded and version controlled within the work, rather than saved elsewhere.

These working practices achieve a range of important outcomes. Minimising manual steps makes analyses faster to execute. This makes it easier to deliver timely outputs, dashboards and reports that reflect the current raw data, rather than out of date information. This speed and low cost also makes it easier to re-execute the whole pipeline swiftly when errors or shortcomings in one aspect of the work are found and addressed, or when modifications have been implemented. Sharing code widely allows others to see the work, and to re-use it in their own identical or related analyses where helpful. Open code also adds an extra layer of assurance, as it allows a wider community of engaged users and experts to help to identify problems, or offer improvements; it also helps build capacity across the system, because people using data can see what others have done with it, and learn from their prior work. Adequate documentation - embedded alongside the code itself - makes the work intelligible to others, and to the same user when they return to the task after a long time on other projects. Formal code review helps to identify and therefore minimise

error, increasing quality, and ensuring trust. Open tools such as R and Python ensure that analysts are not hindered in reproducing, re-implementing or understanding each other's work; it means that entrants from other sectors are able to use familiar standard approaches seen across multiple sectors, making recruitment and onboarding easier; and it means that analysts have access to a vast global knowledge base around those tools.

"There is great work across government to use the principles of Reproducible Analytical Pipelines in data analysis. But this is not yet the norm. Changes are needed if this approach is to become the first choice. This will require collaboration and strong leadership within organisations and across government. This is especially important for health and care data, which is characterised by large data sets but dispersed expertise and limited knowledge sharing."

- Ed Humpherson, Director General, OSR

The principles of RAP are excellent, well thought through, and reflect a strong basic minimum standard. They have not been created in isolation from real practical work, and they are not aspirational: they are widely adopted, and reflect common practice in a range of sectors that use data effectively, and increasingly across government. Some individuals working with data or developing software would regard the core minimum RAP principles to be so self-evident that they barely warrant re-stating.

Researchers and analysts as coders

Beyond the minimal criteria of RAP set out above there are further changes in culture and working practices - towards a more computational approach - which have become the norm in many other areas of academic research, and data science in other sectors. Most if not all of these working practices are also prominent in the RAP approach, outside the minimum requirements set out above. At their core is a recognition of the need for those working with data to embrace the many norms and behaviours of the software development community when writing code. This is crucial, because writing code for data management and analysis in health presents all the same challenges as code for other areas. Any reluctance to recognise this is, in part, an expression of the phenomenon that has held back progress in many areas of health technology: "health exceptionalism", inappropriately overstating the uniqueness of challenges in healthcare, to avoid using effective approaches from other sectors, for reasons such as inertia, or local organisational politics.

"So much of science is computational now. It doesn't make sense to treat infrastructure and research questions as separate. We should also embrace the concept of product management. If we could make NHS [data analysis] a 'great digital product' then you could change the dynamic completely."

- Interviewee

Compared to other sectors, health data presents a more pressing need for a clean, systematic approach to code and data, because of the complexity and interdependencies inherent in the work. The full journey from raw NHS data to a finished analysis or dashboard is typically a long chain of activity spread across many different individuals, teams, and organisations. Often these interdependencies are barely visible, as people simply use and re-use data, code, or other information that was prepared by someone elsewhere in the system.

There is undoubtedly a place for traditional and less technical approaches to collaboration and information management such as emails, meetings, phone calls, or written manuals. But these are ill-suited to managing long complex interdependent technical work with code and data. The software development community has therefore developed a range of behaviours and tools to manage such collaboration. Some of these norms are technical, but it is useful to give a brief overview, in order to lay out the tangible nature of better working methods, and to shed light on the challenges around adoption. Readers familiar with these approaches may wish to skip this section.

Version Control and GitHub

Version control is the process of tracking and managing a project's code throughout its development. Version control software keeps track of all changes made to the code: it allows multiple researchers to work on the same code at the same time, propose changes, allow those changes to be reviewed, and then merge all their changes back into one "main" codebase, and keep an organised audit of all this work. It also provides a safety net, as code can easily be reverted to an earlier version if a problem is encountered later in the project. Working with tools like Git and GitHub to do version control is at the core of collaboration on code, and a range of tools have been developed over time to check for conflicts when changes are merged, run tests on code automatically, and so on. Typically, people share code in the same place that they developed it, usually GitHub. This means everyone can see the history of the code, and the discussions that led to certain design choices, in situ.

Code Review

Code will often contain shortcomings, or errors. Code review typically involves a separate person examining the code, sometimes running it. It might be done on a single "pull request" for a change, or on larger pieces of work. It aims to address errors, and provide feedback or suggested amendments to improve efficiency and readability. One single incorrect character may have a catastrophic impact on an analysis, and this has led to numerous retractions or corrections in academic papers. Many coding errors go unnoticed. Code review is the norm throughout the software development community, and there is extensive training and guidance - with diverse schools of thought - on how to do it well. Open sharing of code makes review easier.

Functions

Often the same task is performed many times over in a given analysis, or across projects. Inexperienced coders might copy and paste code "patterns", with minor changes, for these repetitive tasks. More experienced programmers build reusable "functions", which group the repetitive tasks together into single units of code with their own associated documentation. This can make work more efficient within a team, providing solutions to common computational problems, and makes it easier to share performant code to others. Using functions also helps minimise error, as changes are made once within a function, and this re-used function is more likely to be well reviewed.

Unit tests

Unit tests, alongside code review, are another way to minimise error. A coder writes a unit test to provide a function with a range of controlled inputs, and the expected outputs these should generate; the unit testing framework will raise the alarm when a change or improvement to the code causes an unexpected change in functionality. Good unit tests also provide invalid inputs, and check that the function deals with this in an appropriate way (for example, by raising an error).

Libraries

Useful functions often outgrow individual projects and build a broader user-base, especially when a large number of users are all trying to solve the same suite of related problems, with a range of related functions. When this happens, more experienced programmers move the work into reusable code "libraries" and share them through package indexes or archive networks. The process of creating and sharing libraries can improve the quality of code, because work that is more widely used is likely to be more widely reviewed. Popular libraries tend to be well documented and come with clear explanations and examples, which decrease the barriers to entry for inexperienced coders: when more people use the work, more people invest in improving it. By creating and sharing a library, researchers contribute to the broader research community. This more advanced variety of code sharing is common in many areas of scientific research, such as Geographic Information Science, but it is less common at present in health data research.

Documentation

Analytic scripts can be long and complex, functions and libraries more so: good documentation can improve reusability and understanding by providing information about what each section of the scripts is doing, and why. Where code is intelligible to others, it is more likely to be used and improved; it is also more intelligible to the original team, who may not return to amend or extend code themselves for many months or years. It is important to draw a distinction between well documented code and other forms of documentation, such as static "user manuals", Standard Operating Procedures, or comms material that describes or celebrates a project. The simplest form of documentation is as a "comment" or text note in-line with the code, giving plain-English descriptions, justifications or context for the adjacent commands. Functions will have more formal documentation, and again

there are extensive norms and tools around this. For example, Python has “docstrings” that can be used to describe how a particular block of code can be invoked and used, alongside its expected inputs and outputs; various documentation tools can then pull the contents of these docstrings into a longer more formal manual. Most code - but especially libraries and functions - will also have some overarching contextual documentation with a description of the project, and so on. Good documentation is so foundational to software development that most code management tools prompt users to create a “readme” file with basic information about the work at the commencement of every new project, by default.

Managing the environment

This is the most complex concept covered here, but it is important to help understand the modular and interdependent nature of work with data in the modern era. It is almost impossible to conceive of an analysis script or project that exists in isolation: all work relies on pre-existing libraries or resources made by others, whether these are functions or libraries for arithmetic, data management, statistical tests, data visualisation, or underlying features of the computational environment in which the code is executing. All code makes assumptions about those resources. But functions and libraries are constantly evolving and advancing, typically for the better; as a consequence, commands that once worked in a certain way may have changed their implementation, default parameters, or have been removed or replaced entirely. At best this can prevent code from running; at worst, code will run, but deliver incorrect outputs without the user realising it. By managing and cataloguing their environment, the external dependencies on which they rely, people working with code can avoid these problems.

These working practices are listed not because every reader should implement them in their own work (of course), but rather to illustrate the true mechanical nature of the work; to provide a benchmark against which to judge claims that computational skills are already the norm; and to demonstrate that open working methods and code-sharing are a core mechanical and practical feature of delivering good, re-usable code in a rich ecosystem.

The Limits of RAP and Computational Working

Here the limits of these working methods are briefly considered.

Who needs these skills?

It is not necessary for every individual working with data at every level to develop advanced skills in computational data science: for example, there are vitally important roles for individuals with good skills around interpreting data, and communicating it effectively to clinicians or managers in the NHS in order to help them effect change. There is also an important role for individuals working as analysts, using more “point and click” interactive tools developed for them by others who have deep skills in data management and analysis. It is, however, crucial that a large number of people have basic skills in this space, where they are developing and implementing analyses; and it is crucial that the core working practices, such as sharing code, are implemented as a norm throughout the system, because of the problems that closed working can create around quality, safety, usability, credibility, and review.

Examples of withheld code in NHS service analytics

One senior member of a prominent NHS analytics team described a situation where they were contacted and instructed to validate - or sense check - work that had been done by another team to identify the number of patients with a given set of conditions in a series of geographical areas - a ‘segment The initial work had been done at substantial cost by a commercial vendor of data services (who can often provide good outputs). “Analysts are being asked to respond to segmentation information without anybody being able to replicate it. For the analysts this immediately makes it useless. The analysts cannot replicate the work or see how it works. If the company is producing something truly exceptional in this space, then that is great, but we very much doubt it. It is simply some codelists and date ranges. The NHS should not buy blackbox analysis ‘full stop.’ They should say, if we pay, you have to share the code, we have to at least be able to see it, evaluate it, everything the NHS buys should be open for... scrutiny.”

Another NHS service analyst described a major nationally funded project aiming to produce dashboards of local service activity in primary care. Despite this work being nationally funded, and published in academic journals, the analyst found that the codelists and logic used to identify patients in different categories for the dashboards - which could have been re-used elsewhere - were unavailable for re-use, or for evaluation and checking.

Point and click

There is a legitimate and widespread ambition to have point and click tools for analytics in the NHS as well. This is reasonable and achievable. However, it can only be realised by teams that understand the true underpinning reality of how patient data is generated, extracted, prepared and used across the system; who have RAP and computational skills; and who can work alongside people with clinical analytic needs to iteratively develop prototypes of interactive tools; and can then work with other tools to harden the most suitable of those prototypes into scalable interactive tools. Any solution that appears to not entail this kind of process, and workforce, is simply hiding the work, by commissioning it in less effective and closed means, or similar activity. By doing so, the system is prevented from learning about its own data, and developing the open commons of knowledge and workforce that drives high quality analytics and research in other settings and fields.

Build versus Buy

There is a longstanding discussion in the NHS, and less visibly in academia, about where the line should be drawn between what is bought from the market, and what is built bespoke, with regard to digital tools and services. The focus of this chapter is on open working for data management and analysis; but as this work expands into larger packages and libraries it edges into Digital Infrastructure, which in the modern era is best conceived of as “open code and skilled teams” rather than “beige boxes of computer equipment”. Overall, as discussed in the sections below, the best approach is: re-use existing commercial or open tools that already exist for large tasks (databases, etc) according to which is the best, with a preference for open to ensure access across hundreds of organisations in the NHS; procure open code from public and private vendors for new substantial tasks; build in-house and procure open code from public and private vendors for the vast ongoing workload of data management and analysis within those larger tools.

Current working practices in the NHS and academia

During the review the team spoke with leaders and researchers from other academic fields such as structural biology, structural genomics, and physics where these working practices have already been widely adopted. This can be seen in the platforms, the people, the work, and the outputs. On large scale physics experiments, within projects such as the Large Hadron Collider, it would be inconceivable for data management and analysis code to be withheld. Furthermore, contributions to this code, and the infrastructure to implement it, are well recognised in funding, and in authorship or contributorship to papers and other outputs that result from it. As a consequence of this, in some fields it is common to find papers with hundreds of authors: this recognises the reality of "team science" approaches to such work, and an awareness of the deep integration between developing innovative research methods, developing single analyses, and developing the software that underpins both.

In structural genomics the norm was established two decades ago that all genomic data would be shared openly with the wider community: this was a hard-won battle, driven by a number of senior and junior scientists. However, the establishment of this norm helped drive forward many other aspects of open working across the field: where shared data is being accessed in shared open resources, then it is more natural to share the code that executes against it.

It would be wrong to say that these working practices are never seen in health data research. There are strong positive exceptions to be applauded and built upon. Some aspects of modern open working practices are doubtless exhibited within groups, but without sharing the code (as discussed separately below). Overall, however, it is fair to say that these practices are far from the norm in any of the three key fields that will improve patients' lives through data: academic research, NHS service analytics,

and development of secure analytic platforms. From interviews and desk research it seems that such working practices are more common in groups with strong crossover into adjacent fields where open and software-driven work is more commonplace; and that there are often isolated projects (such as some packages around GP data management) that are created and shared but not sustained, likely for reasons beyond the control of the staff involved.

The team conducted a rapid informal overview of the GitHub repositories of major organisations and recent publications: overall it is common to find that outputs and projects from major organisations that were delivered through code do not share that code, giving only narrative descriptions of the data management and analysis in free text (including for prominent projects from organisations and teams that have spoken publicly of their support for open working, and that are regarded as being at the cutting edge of research with health data); and, with certain very good exceptions, it is unusual to find the robust computational approaches described above used as a matter of routine. For example, it is uncommon to find well documented libraries of performant code (with impressive exceptions). Of note - reflecting a system at the early stages of a transition - there were examples of GitHub being used as a "copy and paste" archive to share code, or a subset of code, at the end of a project (which is not its strength or purpose); GitHub being used as a place to share only free text descriptions of work that was then described as "being on GitHub"; and similar activity. Related, the team conduced a rapid overview of code sharing on GitHub for analytic work from key NHS service analytics organisations of varying sizes including Commissioning Support Units, Academic Health Science Networks, and national and local NHS service analyst groups. Again, with the caveat that there are some outstanding positive examples of excellent practice, it is uncommon to find evidence of routine implementation of RAP and other contemporary methods described above. In interviews the team discussed whether these methods are being used but in non-shared forums: again, with pockets of excellence, there was no good evidence that this was the case. It may be useful for others to repeat this work in more detail.

Barriers to RAP and computational working

The team discussed adoption of modern open computational working practices with a wide range of senior and junior researchers and analysts, across a range of skillsets. Many individuals expressed a strong desire to work in this way, and new entrants from adjacent fields expressed frustration and surprise that it was not facilitated. A range of obstructions were identified, broadly under five categories: skills, recruitment, platforms, funding, and recognition.

Skills

It is clear that there is a shortage of skills and training around these working practices, a widespread desire to access such training, but also a need to have it recognised by senior leaders. There was an awareness that there is an almost limitless array of self-directed online teaching through services such as Coursera, or some MOOCs (Massive Open Online Courses), but no clear signposting or curation of "journeys" through these courses, or guidance on which to choose. Related to this there are challenges around finding work of the right level: for example, there are excellent resources from the Turing Institute around reproducible analytics, but these are very detailed and to a degree assume extensive resource and support to address that issue alone.

Each workforce expressed somewhat different concerns. For academics, there was a sense that more traditional in-person training in this space could often tend towards more generic courses on data analysis in tools such as R or Stata; and that courses focused more on computational methods were very closely tied to delivering analytic outputs in other scientific fields with a stronger tradition in bioinformatics, reflecting the source of the courses, rather than analysis of NHS electronic health records or similar datasets in healthcare. There was discussion of a very recent round of UKRI investment in training in this space, where most outputs are not yet due, but which may help address some of these issues. For NHS service analysts - as discussed in the chapter on this crucial workforce - there is less access to formal training. The recent development of some online discussion forums aiming to drive further training was welcomed, but at present this does seem somewhat limited to links out to various YouTube videos on analysis in adjacent analytic fields, without clear curation of quality, appropriateness, priorities or journeys. For both groups there was concern that - because these areas of work are often regarded as lower status, where the practical aspects are less visible to senior leaders - it could be hard to get permission for access to time or resource for training, or for that training to be subsequently recognised as denoting valuable new skills.

Recruitment

Related to this, there were widespread concerns around inward recruitment of individuals with strong skills in software development and data science for a range of different reasons. Firstly, salary levels were widely regarded as unrealistic, given the very high salaries that those with such

skills can expect as day-rate contractors in the public or private sector. Very senior leaders in the academic sector expressed deep frustration that they had been unable to persuade their own universities to pay developers at anything approaching a realistic rate, citing unrealistic pay scales that "tie pay scales to how many staff you manage"; assume analysts are a low-level "implementation" workforce; or view such staff as being in a similar category and pay-band to those in IT support teams, apparently without recognising the diversity of roles. Similar concerns were raised by NHS analysts.

"Universities conflate research data engineers with the IT staff that run the systems in the university. This needs to be fixed along with adequate support from funding.....They're hard to get and hard to keep. I keep them in spite of my university."

- Interviewee

Partly as a consequence of this problem, there has been a tendency to recruit coders and data scientists from adjacent fields, which can create challenges when their skills or career ambitions are not necessarily aligned to those of the team. For example, computer scientists from university departments may - like funders of computer science work - be more focused on innovations around abstract principles in computer science, rather than Research Software Engineering. Similarly, a new arrival with excellent data science skills from working in bioinformatics, or solar physics - and therefore more willing to work on a lower grade research salary - may not initially be able to engage well with the challenges of a team working on health data.

Related to this, it is clear that there are challenges around training to on-board those arriving, from a generalist software developer or data scientist background, into work on NHS service analytics or academia. Where they are part of a team, this may be regarded as unnecessary, in the belief that others in the team have that knowledge (which may compromise the ability of teams to work creatively to develop analyses and tools, as this is best done when skills are somewhat pooled and overlapping). Where developers are expensive staff on a day rate, there seems to be a view that paying for them to be trained in the basics of epidemiology or other aspects of NHS work would be an inappropriate use of their expensive time. As a consequence, it is uncommon to find experienced software developers with industry standard skills who also have strong domain knowledge across topics such as: the nature of NHS data; how research is done with electronic health records or related research data; the clinical context for such work; the operational context of the NHS; how data is collected, manipulated, and extracted from clinical systems; and so on. This is particularly concerning given that senior individuals in the NHS and academia repeatedly expressed frustration that they wanted to recruit these staff, but that this is "like hunting unicorns".

The RSE movement started principally in the UK but is now international with many national groups around the world.

Lastly, developers expressed concern about the appeal of working in academia or the NHS due to a variety of current working practices. For example, while some developers will accept a lower salary for public service, many equate this "public good" with contributing to open source sharing of code, whereas in academia and the NHS code is often closed. Related to this, developers typically use evidence of prior work to get each new job, and often use their GitHub activity to help future employers see their productivity, and quality; this is not possible when code is withheld from open view; when projects are slow to deliver; or where the software contributions are low status or hidden. Lastly many developers regard themselves - rightly - as high-status team members, which does not sit well with some current assumptions that software contributions are low status, low salary, and "just implementation".

At the end of this chapter, there are range of recommendations to address some of these challenges. In addition, excellent work has already been done to address these problems in other fields, in particular by the Research Software Engineering community (box). There seemed to be limited awareness of this community and its organisational structures among those working with health data in the NHS and academia. This represents a substantial opportunity to work with - and expand - an existing framework to address current challenges.

Society of Research Software Engineering

The Research Software Engineering community has developed over the past decade and celebrated its 10-year anniversary in March 2022. The RSE movement started principally in the UK but is now international with many national groups around the world. The Society of Research Software Engineering was founded in 2019 to help drive recognition and impact for those working in this field: "Our mission is to establish a research environment that recognises the vital role of software in research. We work to increase software skills across everyone in research, to promote collaboration between researchers and software experts, and to support the creation of an academic career path for Research Software Engineers."

This community have done excellent work, and there are now many Research Software Engineering teams in universities across the country, as well as many more individuals directly embedded in research groups,focused on improving the quality, re-usability, and sustainability of research software created by various projects in diverse departments.

Effective RSE groups tend not to focus on theoretical aspects of computer science - although they are informed and trained in these principles - but rather on close collaborative delivery by experienced software developers who also have knowledge from a specific discipline. As such, they are not a separate group of remote individuals delivering "implementation" to a commissioned set of instructions, but contribute at every stage from development of ideas through to all outputs.

Platforms

There was substantial concern expressed by those with skills in open approaches to computational data science that they were actively blocked by the platforms and tools available to them in both the NHS and academia. On a small scale, at the level of individual laptops, they were often concerned to find that local IT policies actively prohibited the installation of tools - such as R, Git, Python or Docker - which they viewed as being a basic minimum necessity for the delivery of their work. Where it was possible to push through these obstructions, it took very substantial effort on the part of individuals, often requiring substantial local organisational and administrative support from senior leaders, over a very long period of time. This has also been an issue in other areas of government, during the adoption of RAP working practices: the team was told of very senior support being needed from the centre of government to get some analysts access to Python for data science work.

It is reasonable for local IT teams to be cautious, especially for more versatile tools such as Python, and especially when staff are working with more sensitive data. This is one reason why it is better for this kind of work to be done in a small number of platforms for secure research, as discussed in the section on Trusted Research Environments: when analytic work on sensitive data is done remotely in a platform, it is not necessary for a large number of small teams in a large number of small organisations to each separately deliver risk evaluation and secure installation of a wide range of tools.

Unfortunately, these problems also seem to be prevalent in many of the large and small platforms that have been created for secure analysis of data. For example, numerous analysts expressed concern and surprise that they weren't able to use GitHub inside Trusted Research Environments, or Data Access Environments, because communication with outside resources such as these were locked down for security reasons; and furthermore, that they were not even able to access more closed tools for code management such as GitLab, as they were either unavailable or implemented so poorly as to obstruct their normal use.

Lastly, in a technical environment where all the tooling, working practices and assumptions are built on a model substantially less evolved than RAP, procurement and implementation decisions are built around an assumption of people using point-and-click and copy-and-paste methods, rather than scripts, meaning that individuals with computational skills simply cannot use them.

Funding

Funding and recognition are closely related, as there is a practical relationship between the two. The challenges here are captured well in a paper from the Wellcome Trust Data for Science and Health team in 2021:

"Unfortunately, the academic funding model wasn't conceived to support software-based tools or to maintain digital infrastructure, and thus clearly needs substantial changes to acknowledge the computing-heavy nature of modern research... Research relies on software... and yet, the systems in place to credit and support those who write the code and build the tool are insufficient. To make research software sustainable, we must adapt our credit and reward system, and ensure that we treat software as not only something that underpins research, but also as a first-class output. It needs to be funded, maintained and have viable career paths even if the researchers involved are writing more lines of computer code than lines in an academic manuscript. Researchers shouldn't have to choose between producing reproducible high-quality code and career progression. Moreover, the risk... doesn't only stem from continuing to disincentivize a necessary part of the modern scientific endeavour (that is, software development and coding), but rather from having vast swathes of the scientific literature potentially becoming even more unreproducible as the [code] infrastructure that made the analysis possible falls into disrepair... Until we collectively acknowledge the need for better support structures around computational research, the status quo will persist."

- Knowles, Mateen, Yehudi, 2021

Numerous individuals complained that they were able to find no open competitive or conventional sources of funding to support them as individuals, or as a team, to focus on the software and related methodological aspects of delivering high quality outputs from NHS data, and that their salaries were therefore only covered as a component part of a grant focused on delivering a specific research paper output, adding to their sense of lower status, and obstructions in both their career and their ability to innovate, as they were only employed, conceived of, tasked, and supervised as "support staff" for traditional academic epidemiology research skills. In desk research it proved extremely difficult to find open competitive sources of funding for this kind of work from any national funders, whether project-based, or person-based.

More concerningly, a number of individuals described in detail situations where a substantial public investment had been made to deliver work closer to research data infrastructure, code, and methods, but that this resource had been - in their view - diverted onto delivery of traditional research outputs, and staff with only skills to deliver those outputs, either because the specific piece of funding had been administered and awarded from funders to individuals with traditional research paper skills, rather than those with computational skills, or because those in senior leadership and strategic roles in their organisations tended to be those with a traditional focus on single research paper outputs rather than code.

"What happened was that a good chunk of that money ended up going to clinicians PAs, research funding, postdocs, and whatever else. They added value in a sense because you really need domain experts to understand the data once you get it out of the system but we always felt that the funding that we had should go on infrastructure because the researchers were already funded elsewhere, it didn't work out that way.....Leadership by epidemiologists has been really problematic for us. They are brilliant at getting research grants and, to some extent that's fine because that's the money we've been reliant on, but they are invested in the existing system. They don't have any particular interest in... infrastructure as a priority."

- Interviewee

Recognition

Researchers and analysts with strong skills in RAP and computational approaches to data science expressed a range of frustrations that clearly reflect fixable structural challenges. At an individual and organisational level, there was an impression that these skills are under-valued, or that they were viewed as “just programmers”. In academia examples were shared of developers and data scientists not being named as authors on publications, or grants, despite having contributed - in their view - a very substantial amount of the work, including creative and innovative methodological work to deliver the outputs. Examples were also shared of situations where individuals had left the field and moved into other industries where they could get both higher salaries and better recognition for their skills.

Some academics express the view - less often in public - that contributions to code are less creative, less intellectual, and “less scientific” than contributions on statistical methods or study design; and that the community should develop some other metric for contributions on code, infrastructure, and similar activities. The concern here is two-fold. Firstly, there is no other metric, and the team found no evidence of any substantial activity to develop any new metric: publication authorship is the norm; it is used for similar diverse contributions in other fields; and it could be used here. Secondly, the sentiments may not reflect the true nature of scientific research. There is no clear reason why the single choice of specific statistical model used to evaluate an association between two variables in an analysis - or the development of an over-arching clinical question - should be regarded as expressing any higher variety of creativity, technical knowledge, domain knowledge, or uniqueness of thought than the myriad highly informed and complex design choices made at every level of the computational work needed to deliver a completed analytic pipeline. Furthermore, the very creative conception of a project - and the rejection of infeasible projects - already requires a deep technical knowledge of the data, the tools, and the extent to which those tools could be feasibly extended, all of which is built on a deep knowledge of the codebase, the data, the possibilities, the research context, the diverse analytic options, and the clinical context. The same is true of innumerable subsequent choices around the delivery of a project from conception to completion in complex NHS data, all relying on deep technical skills around data science and code development, alongside domain knowledge on epidemiology, medicine, NHS operations, and so on. Relegating the contribution of software developers and data scientists in this work to “just implementation” seems unambitious, unrealistic, and unlikely to deliver high quality or efficient research.

Individuals reported similar challenges around other forms of recognition. The Research Excellence Framework (REF) is a large and expensive national programme whereby all universities develop, over the course of years, a complex series of documents and formal filings to describe individual researchers, papers, projects, outputs and environments. These are then used to evaluate the quality and impact of work in each section of each university, a process which inevitably - and in some respects by design - helps to shape institutions’ priorities. While there are theoretically places where software skills and outputs such as GitHub repositories can be recognised in the REF, in practice this is uncommon for work on health data.

As a consequence of these barriers to funding and recognition individuals with computational skills report finding that they struggle to achieve seniority in their organisation. It has proven hard to find many examples of software developers, or those with similar skills, working in this field at a senior level where they are able to shape programmes of work, teaching, or organisational priorities; in some cases, there were individuals with job titles implying such skills, who in reality had more conventional and non-computational research skills. As a consequence the work, software libraries, projects and infrastructure from individuals with computational skills struggle to achieve independent status or sustainability.

None of this should take away from the fact that the UK has an extremely strong, high quality, influential and productive traditional research community in health data research, with a very diverse range of extremely gifted individuals producing a range of traditional high impact research paper outputs on single analytic questions, of great importance to science, which make a substantial contribution to improving patients’ lives nationally and globally, all supported by a robust funding regime that offers a diverse range of open competitive funding for individuals, projects, and larger organisations.

The concern is only about the neglect of the skills base and infrastructure needed to capitalise better on this existing work. Overall, the impression was of a system for funding, recognition, and dissemination that was designed in a pre-computational era, and that is yet to catch up with the contemporary reality of how analytic ideas are iteratively developed and acted on in large mixed teams working across complex interdependent tools, skills, and data. Overall, this should be a source of optimism: there is a pent-up supply of skills; and a strong need for those skills to be unlocked and put to good use. At the end of this chapter are a scaled range of interventions that can rapidly deliver good progress on training, skills, infrastructure, recognition, and funding, including specific funding for these types of work, and light touch oversight that can ensure funding earmarked for these purposes can consistently reach those with the right skills. Beforehand, as a subset of modern methods for software-driven analytics, it is useful to cover the specific challenges presented by open working methods for health data.

The specific challenge of open working

The previous section covered modern computational approaches to analysis and research. The following section addresses a range of specific challenges around open working, the limits of open code, and the strong positive relationship between open working methods and innovative commercial activity in health data science.

It is clear from the minimum working practices of RAP, and the descriptions of more advanced computational approaches, that code sharing is a core feature of delivering high quality, sustainable outputs and data infrastructure.

The practical value of openness is covered above. In outline, open code helps to drive quality through review that identify errors, and by ensuring all users are fully aware of the operations on which they are dependent. It supports efficient re-use, and iterative improvement, in a modular collaborative ecosystem. It supports capacity building, through easy access to prior related technical work. Open code also helps to build trust in statistics from the public, policymakers and professionals, by sharing a comprehensive description of how the raw data was converted into the final analytic outputs to be acted upon; this may be particularly important on contentious issues around performance monitoring, or the risks and benefits of particular treatments, especially in settings where legitimate confidentiality concerns mean that the underlying patient data for a given finding cannot safely be shared.

The Turing Way

The Turing Way is a collaboratively developed handbook, and associated community of practice, that aims to make reproducible data science 'too easy not to do.' It covers all aspects of reproducibility from data management, library sciences, through to software development - including providing training material on specific techniques such as version control - and currently comprises five guides:

- Guide to Reproducible Research
- Guide to Ethical Research
- Guide to Project Design
- Guide for Communication
- Guide for Collaboration

The handbook can be accessed online here and the GitHub repository - which has more than 250 contributors - can be accessed here.

Open Working and Research

The importance of open working is particularly salient for scientific research using data, as the entire scientific process is built on the principle of openness. Researchers do not assert that something is true: rather they detail the methods and results of their work so that others can review it, evaluate it, critique it, and interpret it. This is captured in the latin motto of the Royal Society, from 1660 ("Nullius In Verba") for which the Society provides an official translation: "take nobody's word for it". The complete code for the complete analytic pathway inherently contains a complete and unambiguous description of how the work was done.

This is less about trust, and more about ensuring that the outputs of research are valid. It is common to find that the same general type of analysis, on the same clinical or operational question, in the same general type of population, can deliver quite different answers, sometimes to a striking extent. There are many examples of this in the scientific literature. For example, various different analyses reported different findings for the relationship between ethnicity and risk of COVID-19 infection, admission, or death. The reasons for these kinds of differences can be hard to explain: it might be differences in the source population (for example, one set of GP practices in one research database, and a different overlapping set in another); it might be differences in how the data was managed (for example, different approaches to converting the raw GP records into analysis-ready datasets, as discussed in the Data Curation chapter); or differences in the specific statistical model used to evaluate the relationship between a given patient feature and a given clinical outcome. This is not strategic, informative diversity of approach, because closed analytic pipelines make it hard to know how the analyses differed.

These differences - especially when they are on important topics - can be a deep source of frustration and confusion for clinicians and policymakers, and resolving them is important. To be clear, different groups will often have good reasons for choosing different approaches to different aspects of the work, and this is not about understanding or asserting who is right or wrong. It is simply about being able to see what people have done, seeing the differences, understanding them, and understanding how sensitive the finding is to each difference. Unfortunately, code is commonly left unshared in epidemiology in general but is commonly shared in other disciplines.

The prevalence of code sharing

The team conducted a rapid review of code (during the summer of 2021) sharing for reports created during the COVID-19 pandemic; this should not be regarded as criticising organisations, in the context of ongoing culture shift; and many of these organisations have delivered impressively towards open sharing in other settings. Overall, it may be helpful for others to do more formal or regular overviews: however any audits of open code should be handled sensitively and with a positive focus on helping platforms, organisations and teams to improve standards.

As at Summer 2021:

- **ONS covid reports:** the team was unable to find any analytic code for the platform or covid analyses (but extensive and excellent open code training elsewhere).
- **OpenSAFELY covid reports:** all code for the platform, data management and analysis all shared automatically on GitHub Declaration of Interest (declaration of interest: BG is PI on OpenSAFELY).
- **PHE covid reports:** the team was unable to find any analytic code for PHE reports on topics such as ethnicity and COVID-19; but extensive code sharing for their (excellent) COVID-19 dashboards.
- **DECOVID (Turing / HDRUK PIONEER platform created for a wide range of covid research teams from a large number of universities):** the team was unable to find code for the platform or analyses.
- **ICODA (HDRUK's flagship COVID-19 data analysis platform initiated in June 2020):** the team was unable to find code for the platform or analyses (but also no outputs to date).
- **HDRUK / NHSD / BHF TRE:** the team was unable to find code for the platform; but some scripts are shared for a paper describing the data accessible through it, and one research preprint (the platform's only output to date).

Barriers to open working

There are many strong examples of code sharing among those working with health data, but it is clear that this has not yet become the norm among academics, NHS service analysts, or those building and maintaining health data infrastructure such as Data Access Environments or Trusted Research Environments. Over the course of interviews and desk research a range of barriers became apparent. A summary of these is given below.

Skills and knowledge

As above, researchers and leaders may not know what code sharing is, why it's important, or how to do it; there is little guidance on how to share code informatively, how to avoid unhelpful "code dumps", and how to annotate quickly but adequately; those who work exclusively with "point and click" tools may have no code to share.

Anxiety

Researchers may be anxious about others using their code to question their methods (although this can happen regardless of code sharing); or anxious about sharing imperfect and poorly documented pragmatic analysis scripts (although sharing adequate code is valuable); or anxious that sharing code before publication may compromise journal acceptance.

Lack of obligation

Although some funders or TREs suggest that code sharing is important, it is rare for this to be checked or enforced. Many researchers feel there is no expectation on them to share code.

Lack of resource

Good code management should ideally have been done in platforms such as GitHub where sharing is simple: where this not the norm, then sharing does bring some additional effort; similarly where TREs actively obstruct the use of Git or code sharing by design (typically by de-prioritising code management in their security engineering).

Concern about legal liabilities

One researcher reluctant to share code expressed the view that they could not, because they would have legal liability for any subsequent use of their code, even codelists for an analysis published in an academic journal (for clarity this is reported as a view expressed, but not endorsed).

TRE design

Many TREs have made it difficult or impossible to use tools such as GitLab or GitHub, and make it difficult to share code, as a consequence of these user needs being de-prioritised in their approach to security engineering.

Lack of credit or reward

Researchers and analysts need funding and recognition to make their work sustainable, and feel that their work, if re-used, should deliver some attribution when it is re-used.

Active obstructions in academia

Some researchers aim to preserve a competitive advantage against other research teams through their own private efficient methods for data management; some may have an ambition to commercialise their data management or analysis code, often unrealistically or for trivial revenue; some feel obliged by funders or seniors to "protect new intellectual property". (It is worth noting that code can be shared for review while still retaining rights on, for example, commercial exploitation, in specific circumstances where this is felt to be in the nation's interest).

Obstructions in government

There are legitimate reasons for government to conduct operational research in private, ideally with a plan to share code later; however even under these circumstances it is advantageous to share code between groups, as is easily done with tools such as GitHub or GitLab for smaller groups (although less so for very large networks of users).

Cultural

Knowledge may be divided within organisations (for example, junior analysts may be more aware of current best practice than their senior leaders); and there is a risk of "the best being the enemy of the good" (some organisations and people have advocated perfect approaches to code sharing that are unattainable for many analysts).

Pseudo-open working

As a consequence of growing support for open working, there are now individuals and organisations who state that they support open methods, but do not do so; or create only the appearance of open working. During this review the team encountered examples of very senior and influential leaders extolling the virtues of open working, where their published papers from the pandemic in 2021 do not contain code, and require that interested parties contact them personally to negotiate access to the data dictionary codelists used to define the variables used in the analysis. Similarly, examples were encountered of prominent organisations that use GitHub as a place to store some free text information, and then state the "the project is shared on GitHub". Overall, such phenomena can be viewed as a positive sign that the cultural norms are transitioning towards more open working.

Most of these barriers lend themselves readily to very obvious mitigation. In addition, the majority of academic code is not withheld for any kind of strategic or personal objective, it is simply an unhelpful habit that has emerged over time. There is a clear need for training and incentives around modern, open approaches to data analysis, as discussed above, with interventions set out at the end of this section. Five of the barriers described in this section warrant further exploration.

TREs that obstruct code sharing

TREs have typically been designed to obstruct interaction with external resources, as part of delivering a secure environment for analysis of sensitive patient data. As a consequence, many make it challenging to work with services such as GitHub to develop, manage, and share code with colleagues or others. This is readily surmountable, where the user need is prioritised. This is not an unreasonable issue for TRE teams to consider: sometimes, in certain situations, code developed iteratively against real personal data may - unintentionally - contain a small amount of information that might be regarded as presenting a disclosure risk for information about an individual: while this risk is low, and the nature of any information released is unlikely to be disclosive about an identifiable in normal circumstances, it is always important to remain highly vigilant. Supporting modern, open, collaborative approaches to data management and analysis is an issue that TREs must prioritise as part of their service design: it is usually straightforward to implement internal services for code management such as GitLab, and check code when it leaves the secure environment; while TREs such as OpenSAFELY make it possible to interact freely with GitHub, and indeed require code to be on GitHub before execution.

NHS or government monitoring performance in private

Open code does not mean that all data is shared. Detailed and disclosive patient data should never be shared openly, as discussed in a later chapter. Open working does not also require that all aggregate data created or used in the system is shared as open data. Making the choice to share data as an open data set is a very specific one, made on the basis of operational requirements. There are strong arguments for a presumption towards sharing data openly, even where some in individual organisations have concerns, and this has been a valuable tool to improve performance across the system, when done thoughtfully. However, sharing data is a different issue to sharing code, and the merits of open or closed performance monitoring are outside the scope of this Review. Lastly, it is reasonable for the system to prefer, in certain circumstances, to execute analyses discreetly, for example when conducting initial evaluation of a clinical problem: there is a need for an open policy discussion around what the criteria or limits for this kind of closed working should be, and how long such analyses could or should remain closed.

Legal Liabilities

As above, one researcher expressed strongly to us their view that sharing any code from an analysis, even the list of SNOMED-CT codes used to create a variable in a published academic paper (see Data Curation), would expose them to legal liabilities, for example if their work was taken, modified, and re-used by the manufacturer of a medical device that is regulated by the MHRA. The team raised this with the MHRA: in brief there seems to be some current ambiguity under which a researcher could arguably be interpreted to have produced a medical device by sharing code. In our view this reflects a broader lack of clarity as the MHRA moves to address new challenges faced by the regulator in managing and evaluating digital health technologies. Furthermore, most

medical devices are likely to re-use unchanged or modified versions of huge volumes of general purpose open source functions and libraries from outside the field of health since (as is widely recognised and documented) these are at the core of most commercial products in all sectors; it is inconceivable that the individuals or organisations providing every re-used python or C++ library should be held responsible for all subsequent uses by the MHRA. The review team cannot adjudicate on this issue. It would be useful for the MHRA to clarify for the avoidance of doubt that academics can share codelists and code, on the basis that companies modifying or re-using other peoples' work in a medical device that they themselves market should take responsibility for their own product; or if there is complexity in this space to surface that clearly.

Commercial models and academia

Universities have had a longstanding interest in commercialising intellectual property, and it is right that the nation and its educational institutions should aim to avoid substantial value being captured by others. However, the examples presented to the team during the review for commercialisation specifically of code generated for health data management and analysis tasks represented revenue streams that were either very modest, with no clear paying customer base, or largely implausible. The examples given of academics commercialising code around data science work typically resulted in very modest income for individuals, often at the cost of reduced transparency, reproducibility, and therefore credibility or reliability for the scientific work underpinning it. Examples were also shared of situations where analysts had been asked by their institutions to withhold code on the basis that it might have commercial value, even though those writing the code felt this was unlikely, that the nature of their work had not been well understood by those asking to withhold it, and that they would not have paid others for the same code. Nonetheless there is no doubt that there will be occasions where an individual or team has produced something of true innovative and commercial value, where the benefits of commercialisation may outweigh the very substantial network benefits of open working across the system: while open should be the default, there is a need for an exceptions process.

Unfortunately, university staff in particular expressed the strong view that the default from the institutions they interact with is that code cannot easily be shared. The team was frequently told that some aspects of the standard contracts from large research funders required or encouraged recipients to seek to own their intellectual property, document it, exploit it, and report on their success in doing so; and that their departments or commercialisation teams in their university would encourage them to withhold code on the prospect of commercialisation. Whereas it was possible to overcome these barriers in situations where the team were happy or keen to share, this took effort, was regarded as unusual and problematic, and sometimes entailed conflict.

"[The] biggest pushback [we've had] on creating open [code] for data management, and so on, has been from academics. People are afraid that others will take their data, or take their idea. In other fields, it's the thing that everyone does. In healthcare it's just not in the culture."

- Interviewee

Commercial Vendors and Open Code

Open working for data science does not mean that all aspects of the entire software stack for all work with data must only ever be fully open source, including (for example) all database software, or all word processing and data visualisation tools. There are often good reasons to choose closed commercial tools for some aspects of work, especially where there has been extensive and longstanding prior investment in a tool, as is more common where such tools meet the needs of a very wide range of users, as with Tableau, or commercial SQL server services; and where there are individuals or teams in the system who (often for reasons of skillset or prior design choices, but also on the basis of functionality) have a strong preference to use those tools. However, for the bespoke specific work of managing and analysing NHS data, it is crucial that code is shared as a matter of routine.

The team discussed commercial aspects of code with a range of commercial suppliers of health data tools and services to the public sector. They generally felt that prior intellectual property created by them as an investment should remain their own IP: this is a very reasonable position. On the question of whether IP created to order by a vendor for a research or NHS client should be contracted as delivery of open code, they were overwhelming supportive, with only a very small minority dissenting. It is important to note that, while the code itself is shared for review and re-use, open source software sits very comfortably alongside commercial models. For example, there are many flourishing commercial models around implementation and technical support of open source software. Some companies make some or all of their code open source in order to make other aspects of their commercial services more attractive, or accessible to users, or to spread their approach to solving a given computational task the norm across other platforms. Similarly, developers, software bureaus and large companies are commercially contracted to produce or modify open source software to meet specific users' needs.

This model of procuring individuals and teams to deliver open code bespoke to government's needs - which is then shared as open - is also the norm in the Government Digital Service where it has been successful at improving quality and efficiency of services. It has been likened to the model of building an extension on your house: when you contract a builder to extend your kitchen, and pay them for their work, then at the end of the job the kitchen belongs to you, and not the builder. Overall, the network benefits of modern open working methods - for the NHS, research, and life sciences - are transformative; and the history of closed working methods in this space, which are unusual by the standards of some adjacent disciplines, has held back quality, safety, and productivity. In our strong view all code whose production is paid for by public funds should be shared as open source code

for review, re-use, and iterative improvement under open licenses, such as the MIT Open License; but with a robust and publicly documented exceptions framework using clear prespecified criteria whereby researchers or vendors can request special treatment, where this can be shown to be in the national interest, or provide some other comparable benefit. More detailed recommendations are given at the end of this chapter.

Government Digital Service and Open Code

Government Digital Service (GDS) have long been proponents of open working and coding in the open. “Make things open, it makes things better” has been one of their core design principles for government digital services since 2012 and the code for gov.uk has been openly available for re-use and review almost continuously since it was first launched. This is in keeping with the requirement in the GDS service manual - the standard against which all Government digital projects are assessed - to ‘make source code open and reusable.’

As well as leading by example, GDS blog about the benefits of open working, provide examples of the benefits from across Government, and provide guidance and reassurance on how to share code safely and securely, as well as how to maintain version control. There is also a GDS-run cross-government slack channel which individuals can use to share tips on open working and ask questions from those more experienced in this way of working. NHS organisations can sometimes perceive themselves as being separate from the rest of Government. To a certain extent they are, but this should not preclude NHS analyst teams, or health data researchers, from making use of these - and other - GDS resources on open working.

How Open Code Supports the Life Sciences Sector

There is also a clear market for digital innovators offering digital tools to the NHS which go beyond the foundational work of data management and analysis. It is entirely reasonable for innovators developing IP with their own investment to expect a commercial return. Moreover, the national and global market of such innovators will be very substantially facilitated in being able to develop innovative digital tools by having access to clear, adequately documented code for data management and analysis, as this represents a complete technical picture of the raw digital material in the NHS with which they can innovate and develop tools for the national and global digital health market. Furthermore, potential commercial users of data - such as pharmaceutical companies aiming to use routinely collected health data - will only be able to have a clear picture of the opportunities and operational feasibility of projects when presented with an open ecosystem, with adequate Data Curation code (see Chapter), and adequate TREs (see Chapter). This open ecosystem is the route to drive inward life sciences investment, and build a thriving life sciences community built on NHS data, wherever this is an ambition. The issue of the NHS taking a stake in innovations that depend largely on NHS data for their development is discussed in the chapter on Information Governance.

Open Code is Not Free

When considering the relationship between open code and commercial models, it is crucial to recognise that Open software is not free. The benefits of open code are many: users and commissioners can see it growing during its initial development, and iterative development in the field; all relevant experts collaborators and users can contribute or feedback; and when completed it is owned outright for ready re-use, modification, improvement, and iterative expansion. But someone must pay for it to be created and maintained, just as with closed software.

It is therefore important for the system to understand how to resource open source tools and ensure that teams with impact are sustained. Open licenses mean that formally code can be re-used, critically reviewed, and modified. But in many cases there are good reasons to collaboratively involve and resource the team that originally developed it, especially when modifying or implementing a very substantial codebase: they will likely have developed deep or even unique expertise in the area, having delivered working code; they may also have deep knowledge of the design and implementation choices not taken, and the good reasons why. The original team - whether private or public sector - may also have skills, knowledge and working practices that the system needs to see grow. It is also important, for example, to be thoughtful about “forks”, where a project splits into two separate branches to meet different sets of user needs.

In the open source community globally, especially at the intersection with the public sector and with science, an extensive literature and culture has grown around this question. There are no hard and fast rules that can define the best approach in a single paragraph: thoughtful resourcing here speaks to the need for the system to be resourced, and productivity overseen, by those with deep experience and understanding of software development and the nature of productive open source software ecosystems.

Open Code and Accountability

Many people at the senior and junior levels across the system expressed serious frustration to us about services that were claimed to exist, especially in terms of data infrastructure, but that turned out on closer evaluation (where this was possible) to be more present in comms material and PowerPoint slides than in reality. Well-documented open code and functions in libraries are a key way to demonstrate good progress on a project, regardless of the team working on it: it can demonstrate that code has been written, is usable, and has been used.

Restricting code access to those within the NHS

The benefits of modern open working methods are vast, and long overdue in the health space. Nonetheless is important to consider whether it is possible to have a hybrid model, where the benefits of open working methods can co-exist with closed code, for whatever reason this is deemed appealing: either some prospect of commercialisation, or a deeply held desire - for some reason - to prevent those outside the NHS seeing the methods used to generate NHS statistics internally.

In many commercial code development environments code is worked on collaboratively using services such as Git, but only made available to individuals within the organisation (who need to see it, understand it, and iteratively modify it) in a restricted way. This is a deliverable model for an organisation with clear perimeters. The challenge is that for the research community, and NHS service analysts, the group of people working on the code (or in a position to use it, re-use it, evaluate it, and iteratively improve it) work in more than one group, or institution. The NHS is inherently and by deliberate design a diffuse collection of organisations of different sizes and statuses spread across the country, which all have different contractual relationships with each other, or none at all. Furthermore, the NHS - or rather, the myriad different organisations within it - often solicits or needs help from other organisations adjacent to a single specific NHS entity, such as a research group, a thinktank, or a commercial provider of some aspect of NHS data analysis. The academic community is similarly, by design, very diffuse.

Creating the technical and administrative infrastructure needed to manage restricted sharing of code in a closed way between so many dispersed organisations would be an exceptionally large technical undertaking, but also an administrative one, as it would require very extensive tracking of personnel and

permissions across hundreds or even thousands of organisations. It would be extremely expensive, and may not even be achievable. Because the community using and evaluating code is so huge, and so dispersed, the system is effectively left with something akin to a binary choice: accept inefficient and outdated working practices that compromise quality; or embrace open working methods as a norm that is increasingly prevalent across government and adjacent academic fields.

It is also hard - but not impossible - to conceive of a model for a commercial market where public or commercial teams could sell access to use multiple small blocks of code to deliver data management or analysis tasks. This is difficult for a range of reasons: firstly, the individual elements of code are often small, which would require multiple very small payments and create substantial friction around access, especially when small blocks of code cannot easily be evaluated without seeing and using them; secondly, this friction will obstruct market entry by digital innovators into other areas of digital health activity, who will be blocked from seeing and using code during start-up work; thirdly, teams will often conclude - rightly or wrongly - that they would be better off simply re-implementing the code themselves; and fourthly, under the current paradigm of multiple small duplicative and inconsistent data analysis environments for NHS data, code often requires multiple small modifications to be re-implemented in new settings, and so does not come as a simple transferable package. In the section on Trusted and Shared Research Environments we discuss the opportunities to rationalise these duplicative environments, and consider an App Store model where code can be produced and retailed to run in these general purpose environments with a wider user base.

A future for health data analytics

RAP and a "software first" approach to analytics should be energetically adopted and supported throughout health data research and NHS service analytics. These everyday working practices for data analysis from other sectors present a clear, replicable model of how health analytics could and should work for both academic health research and the NHS service analyst community. It is also readily deliverable: the RAP programme, with its attendant training and online resources, shows that working practices can be rapidly modernised in adjacent parts of government; and the practices outlined above only reflect existing norms from adjacent sectors and academic disciplines. Furthermore, it is not necessary for all working practices to change overnight: rather it is preferable to have a range of pioneer projects that can demonstrate the value of these working methods, and act as advocates and examples of good practice from within. Ideally these should be academia, and particularly in the NHS service analyst communities and platform communities,

Alongside the deliverability, it is useful to recognise the importance of this modernisation. Policymakers have long expressed an ambition to create better curated NHS data for rapid work in academia, the life sciences sector, and NHS service analytics, releasing all the value in our deep, longstanding, detailed, and comprehensive electronic health records: this cannot be realised by manually creating catalogues, or taskforces. Data curation at this scale is a computational, data management, and knowledge management challenge. It can only be realised by adopting contemporary working methods and by building well curated libraries of portable code for data management for each underlying dataset and derived variable, with good documentation, as discussed in the section on Data Curation. Similarly, the long-expressed ambitions for better use of data to improve the quality, safety and efficiency of NHS services cannot be realised with the current closed siloes of manual work: they can only be delivered by adopting modern, open, everyday working practices from adjacent sectors. The longstanding ambition to broaden access to data while preserving patient privacy cannot be delivered by creating ever more small, closed, isolated data analysis environments that duplicate risk and obfuscate the technical aspects of the work: they can only be delivered by building a small number of Trusted Research Environments that are built on, and support, modern open approaches to data analysis. Lastly, the longstanding ambition to "separate the data layer from the application layer" can only be achieved in the context of a workforce that uses these modern approaches as second nature.

It is not necessary for every service analyst or academic to develop advanced software development skills. However, it is important that a reasonable number of individuals have these skills, and that code is shared by default. It is likely that this will be best delivered by Data Pioneer teams, selected from teams or individuals in the system who already embody these skills and behaviours, training them, and supporting and expanding their work. Below are a range of specific recommendation that can deliver modernisation of working practices, focused on skills, incentives, and platforms.

It is not necessary for every service analyst or academic to develop advanced software development skills. However it is important that a reasonable number of individuals have these skills, and that code is shared by default.

Recommendations

RAP and open computational approaches to data analysis represent a core over-arching principle that runs throughout all aspects of work with NHS data. This section contains a range of specific recommendations to deliver the workforce and working practices set out above, covering a range of organisations, informed by our discussions across a wide range of stakeholder groups. Funders of research with health data (UKRI, NIHR, the Wellcome Trust and others) have a vital role to play in building a productive, modern, open, collaborative ecosystem for data science using health data: they are able to define norms and incentives through what they fund, and the prior behaviours they reward. Data Controllers have the ability to make data access contingent on basic criteria around open and productive work that is high in quality and free from error. The NHS has the ability to set norms, and lead on high standards around RAP, especially with its strong access to local and national datasets. Lastly, the commissioners of TREs have a responsibility to ensure that their services do not actively block modern open working methods.

Establish clear expectations around RAP and open code for the whole system

Open 1. Create a RAP and Open Code Oversight Group

There is clear evidence of inertia on this issue, and coordinated activity across a range of organisations including funders is required to deliver change. The NHS Transformation Directorate should convene a small group to ensure change happens by commissioning, monitoring, promoting or delivering the initiatives below as appropriate.

Open 2. Create a public policy setting out expectations on open code

Create a public policy setting out expectations on open code for public organisations and individuals to support, with broad brush expectations, links to further technical information, commitments which organisations can be held to, and a clear statement on the limits of open code and the compatibility of open code and commercial models.

Open 3. Make open code a boilerplate feature of all public contracts

Open code sharing should be a required feature of all standard contracts between the NHS and any external provider of code for health data management and analysis; with a similar arrangement for academic funders; and for any university or other body sub-contracting such work.

Open 4. Create an "exceptions framework" whereby publicly funded code can be closed by prior arrangement if this meets NHS and UKplc strategic objectives

Individual researchers, organisations or funders should be able to actively apply for a single project to be closed, under a carefully designed "exceptions" framework, where each exceptional request is evaluated to determine whether this exception meets reasonable national, individual or organisational interests around commercialisation and collaboration, and whether the network damage inflicted by a closed license is justified by another greater public benefit. This expert group should include intellectual property specialists, software developers, researchers, key stakeholders with expertise (such as the Open Data Institute), policy experts and funders.

Open 5. Create an Open Code Ombudsman and Assistance Unit

It is possible, particularly whilst the NHS analytics workforce is in a transition period, that there will be occasions when an analyst or team of analysts in one NHS organisation needs access to code that has not yet been made open by a different NHS organisation - potentially including central Government organisations. There is a need for an independent body that is tasked with dealing with these, and other, disputes. This unit should listen to complaints, and feedback common themes to the relevant policy, funding and commissioning teams so that appropriate guidance can be developed. To make this practical, the unit must have appropriate status and power over even the largest NHS organisations.

Open 6. Assert that publicly funded code is publicly owned: cautiously consider "Crown Copyright for code"

At present decisions about sharing and licensing code are made, at small scale, in huge numbers, across the health and research landscape, often by people who do not understand the impact and implications of these choices for themselves and/or the wider community of data users. This has very substantially blocked code sharing, which should be the norm, and is the norm in many adjacent research specialities. An expert group should be convened to formally consider a new national standard: that public funded code is publicly owned, under a formal license that covers all code produced on public funds; with an expectation that all publicly funded code should be shared under the MIT open license; and exceptions to be decided by prior arrangement with a prespecified ruleset.

Open 7. Data Controllers should require RAP and open code sharing from data users

Data controllers and especially the NHS should require all those accessing patients' data to share all analysis code, or openly argue for exceptions in single cases. Where the system is in a period of transition, this should be strongly considered in all data access requests, and where there is no plan to share code immediately there should be a credible plan to ensure that this is done later. As with other mandates around open code this should have a clear pre-specified exceptions framework.

Open 8. Amend the Code of Practice for the Research Powers of the Digital Economy Act

The Department of Health and Social Care and the UK Health Security Agency should work with the Department of Digital, Culture, Media and Sport to make a minor amendment to the Code of Practice for the Research Powers of the Digital Economy Act (the primary legal gateway for accessing non-health data) requiring data analysts to make their code available, with a robust exceptions framework as discussed elsewhere.

Open 9. Make it "Okay to Ask" about access to publicly funded code

The team heard from several interviewees that at present it is commonly regarded as provocative to ask for access to the code used to implement an analysis, especially in some parts of the academic community, despite general positive statements on open working. Culture change will only be possible if it is deemed socially acceptable to question when deviations from the 'new norm' of open working are identified. This should be made clear in all relevant policies, and codes of conduct, across academic and NHS organisations.

Develop Guidance on Open Code from Specific Key Organisations

Open 10. Health and Care Information Governance Panel guidance should facilitate open code

It is important that all NHS organisations are given clear direction that code sharing is not the same as data sharing and that it is entirely possible to share code routinely and safely without the organisation incurring significant costs or reputation risks. To make this clear, the Health and Care Information Governance Panel should create guidance on the importance (and permissibility) of code sharing, to go on the Information Governance Portal, emphasising that transparency is a crucial means to build public trust and clinical safety.

Open 11. The Information Commissioner's Office should produce guidance to facilitate code sharing

The Information Commissioner's Office (ICO) is currently working to produce new and updated guidance on anonymisation. Alongside this, the ICO should also produce guidance regarding code sharing. This should make clear that sharing code is not a disclosure risk, and that those writing code have a clear responsibility to ensure that their analytic code is not disclosive of any personal information. Ideally this guidance should also make clear that code sharing and the practices associated with Reproducible Analytical Pipelines are an important aspect of good citizenship around data usage. There is also a need for better guidance and training on ensuring that analytic code is not disclosive of any personal information.

Open 12. The Medicines and Healthcare products Regulatory Agency should address code sharing and device regulation

It is not practical or useful that there should be conflict - or perceived conflict - between code sharing and medical device regulations. The Medicines and Healthcare products Regulatory Agency (MHRA) should deliver absolute clarity to reassure all that where they share code for critical review, re-use and improvement or modification that they do not have liabilities around how it is re-used by others, especially where they have given a clear statement that their code is not intended to be used as a medical device, and especially where the code is shared alongside an analytic or research paper output. Without this clarity, researchers are likely to be more cautious than is necessary. More broadly it is clear that there is a general lack of clarity around medical device regulation and code, and in particular a lack of clarity on remit and what constitutes a medical device, which is likely to lead to some code being both inappropriately included and excluded from the definition, potentially exposing patients to harm. The MHRA is working to provide this clarity, add strong support, and encourage the organisation to prioritise this work. There is an important role for regulation in this space, and some organisations sharing codelists that they have assured and produced explicitly for implementation in NHS service in use as a medical device (such as in a popup trigger) would benefit from being able to wear that assurance and regulatory status prominently.

Open 13. Negotiate co-ownership of claimed commercial innovations from NHS data

See Information Governance and Ethics chapter.

Support NHS service analysts to work with RAP and open methods

NHS analysts are a crucial part of the wider analytic community and can lead by example.

Open 14. Write an 'OpenAnalytics Policy for the NHS'

Bring together DHSC, and the NHS Tranformation Directorate to write a policy that makes it clear to all analyst teams across the NHS, and all general managers, that sharing code is not the same as sharing data and that open is the preferred and default method for all analysis conducted using public data and public funding. This policy should set out best practice for using open working methods, for openly sharing code and for writing documentation. It should also cover more complex areas such as licensing and the protection of IP where applicable. It should be kept under regular review to ensure it remains up-to-date and should signpost to further sources of help and advice where necessary. All external procurement of data science services, whether data management or analysis, should require that all code and codelists produced to deliver the work are shared openly by default. Exceptions to this should be rare and explicitly pre-arranged, with clear justification under pre-specified criteria set by the NHS Transformation Directorate and DHSC.

Open 15. Make Open a Standard Contractual Requirement

Government departments should require all those conducting data analysis on their behalf, in-house and as contractors, to share all code, consistent with RAP and the computational methods above, with adequate technical documentation as per RAP criteria. To aid with this Intellectual Property assignment and publication requirements should be laid out in template and framework contracts so that all organisations commissioning or contributing analysis and code to the public sector are held to the same standards. These contractual requirements should be developed in collaboration with members of the research software engineering community to ensure that they are fit for purpose. It should also be borne in mind that all such requirements, policies, and guidance documents are likely to be 'living' and regularly iterated best on the evolving nature of best practice in this field.

Open 16. Commission intermittent open code audits to drive improvement

In collaboration with academics and key organisations such as AphA and the NHS-R community the NHS Transformation Directorate should commission regular code audits of all organisations that have received public funding for health data research or analysis, including funding for the development of intermediate knowledge objects (for example datasets, or TREs). These audits should evaluate adherence to RAP and open computational methods; follow a set methodology; be published openly; and be used for the explicit purpose of improving performance on code sharing, rather than penalising poor performance. Specific criteria should be developed in collaboration with the community but include: all code shared on GitHub or similar; adequate technical documentation embedded within code as per RAP criteria; use of version control and appropriate methods; sharing of non-disclosive open data where possible; support for continuing professional development in work time; whether staff meet job descriptions with training, continuing professional development or other proof. These elements should be split out into overarching themes to inform targeted interventions for improvement. Good performance should be further incentivised, by highlighting best practice examples. The intention should be to identify blockers to sharing and opportunities generated by sharing - specifically high performers should be asked how they adopted open working methods and the benefits they have seen and poor performers should be asked what help they need in order to 'go open' and then provided with the relevant assistance.

Open 17. Establish a technical writing and documentation team for the NHS

Too often the completion of routine local or national NHS analysis tasks, or the reproduction of key technical platforms in separate locations, relies on an "oral tradition" with documentation passed in conversation or email. This is unsustainable, introduces risk and inefficiencies, and impractical in a massively federated system. Technical documentation improves reproducibility and sustainability but writing good documentation is a skill in and of itself. Hiring a central team to train others and write documentation for key technical platforms & tools, including TREs, python libraries, and more, used across the NHS, would greatly facilitate collaboration, and reproducibility.

Build workforce capacity for RAP and modern, open, collaborative working

The system as a whole is held back by a striking shortage of individuals with crossover capability, covering both technical skills (around data science and software development) and domain knowledge (around epidemiology, NHS service analytics, health services research, and the broader NHS and clinical context of technical and analytic work). This will require training for staff at a range of levels, and for a range of personnel. In outline three key programmes are required: software skills for analysts; NHS domain knowledge for developers; and brief training for senior generalist leaders.

Open 18. Create a "Code For Health" training programme for NHS service analysts and academic researchers

This should include advanced computational data science covering RAP, software carpentry, version control, functions, code documentation, and similar. This should be at a range of levels including MOOCs, short courses, and long courses, with strong practical elements. The following principles and working practices are strongly recommended:

Combine NHS and Academia

Although historically NHS service analysts and academic analysts are considered separately, this is a strong strategic opportunity to begin building robust technical bridges between the two: both work on similar data, often with similar methods or tools; and both should ideally work in similar data analysis environments, as discussed in the Trusted and Shared Research Environments chapter.

MOOCs and Practical Work

Traditional teaching and training has a role. However, for scope and access, a priority should be placed on delivering training as a combination of Massive Open Online Courses (MOOCs) and in-person teaching, both developed specifically for this purpose. MOOCs can cover factual content, and with practical code development exercises can be procured outright: they should be made openly available to all. In-person supervision is necessary for a subset of attenders, especially those seeking certification, to supervise practical work, and for marking work to evaluate competences. This will necessitate a per-person fee, which will create a revenue stream for providers, but necessitate a system for unit payment from the NHS which may obstruct NHS analyst attenders given current lower status of this group in some organisations (see NHS Analysts Service chapter): thought should be given to block purchase or training budgets as seen in other parts of the NHS.

Build on Prior Work but Maintain Focus on RAP

This training should build appropriately on various existing resources and expertise including: the work by existing RAP teams; the work by the NHS-R Community; and other work specifically on RAP and computational methods. It should focus specifically on RAP and computational data science techniques as directly applied to working with NHS data. Caution should be taken that resource is not diverted into training on other issues such as general research methods, specific statistical methods (except as specifically embodied of RAP training), or newer methods such as Machine Learning which are useful but different subjects. Similarly this training cannot be delivered by universities rebadging existing training on other topics such as bioinformatics; or by simply linking out to generic data science training resources from other suppliers; albeit that these may be fertile starting points for modification into bespoke training on RAP and computational methods in NHS data.

Open Competitive Procurement

Training should be procured by an open competitive process, amenable to the best of either public or private providers. All training can be commissioned by either UKRI, the NHS, or both. A rapid technical review should be conducted of recent very welcome UKRI investment in this space to evaluate whether it addresses RAP and computational methods as above, and whether outputs are open. There is currently a rich diverse ecosystem of training in many other aspects of analytics, with no concern about overlap between offers, and similarly a diverse range of providers with different focuses should be acceptable here. An emphasis should be placed on finding providers with strong previous track record of using RAP or open computational methods.

Open 19. Create a "Data for NHS Leaders" training programme

For senior leaders and those in adjacent skill groups training should be accessible on the basics of data analysis, RAP and computational methods so that this group can understand the principles and purpose of the work done by those in their team. This should be MOOCs and short courses. Some of the excellent recent work at the Number 10 Data Science Unit (10ds) "data for leaders" programme may be adaptable.

Open 20. Create an "NHS Data for Developers and Data Scientists" training programme

Very senior leaders in the system repeatedly expressed to us the view that individuals with deep developer skills and NHS domain knowledge are extremely hard to recruit, if not impossible. Strong software development and data science skills are developed over many years and often require a specific disposition: it is not realistic to expect that current NHS analysts (with a deep but very different knowledge base) can be trained in such skills to the standard required by the NHS and wider community. New knowledge for generalist Software Developers and Data Scientists should be addressed with health data "Transfer Training" so that individuals with strong technical skills can develop deep NHS domain knowledge. This training should include the basics of: non-communicable and communicable disease epidemiology, focusing on the specific methodological issues that arise when analysing health data; NHS service analytics; the operation of the NHS; the clinical context; how NHS systems collect and store information; and the structures, strengths and weaknesses of NHS EHR data.
This training can likely be adapted from existing teaching and training on these topics. Priority should initially be given on these courses to developers and Research Software Engineers already working on health data. Attention should be paid to resourcing: the NHS and research sector is actively trying to recruit in and train skilled individuals, from some of the most competitive global jobs markets; this may necessitate paying developers for their training time, and the course fees; it is reasonable to adopt a similar approach to for example MBA funding in the Civil Service, where fees and time are paid, with a contractual obligation for subsequent public service. The developer and data science community are used to working with well-documented and open code; therefore insofar as this can be rapidly replicated in the health data space, onboarding times will be greatly attenuated. Fixing this problem will take a short period of years but pay huge dividends for the NHS and the wider ecosystem of innovators.

Make code a central feature of work in universities using health data

Universities, research funders, and associated governing bodies need to recognise that much of health data science now involves academic researchers effectively writing software, and that this is deeply enmeshed with the work of iteratively developing, implementing and evaluating new research methods. University staff need to be able to access training in these skills and be given protected time and resources to make the most of training opportunities. Those already demonstrating best practice in this domain should be recognised and rewarded appropriately. The following specific interventions are recommended.

Open 21. Modify the Research Excellence Framework (REF) to reflect computational work

REF is used to assess research quality at UK higher education institutions and is known to drive university strategy and promotions for staff. Generally evidence of quality is provided in academic papers, with other sources admissible, though principally through other elements of REF such as Impact Case Studies (which are fewer in number, and therefore tend to be very large suites of work). It is theoretically possible to return outputs such as well-curated GitHub repositories, but in practice this is not often done in work using health data. This can be addressed in broad terms (making RSE and code a required or recommended form of proof of a strong research environment in REF guidance); but also with concrete quantitative requirements. For example, given that almost every quantitative research paper will entail the production of analytic code, there should be a requirement that a percentage of quantitative papers are accompanied by a link to an openly accessible GitHub repository or similar open resource containing the code.

Open 22. Embrace Research Software Engineering (RSE) in health data work

The RSE movement has had high impact in other sectors, and warrants strong support across all sectors; in particular it should be strongly supported to expand into work using NHS data.

Open 23. Pay realistic salaries to software developers

Software developers are among the highest paid staff globally, but university pay scales typically do not recognise the skillsets required to be a research software engineer, and often attempt to hire engineers and developers on pay-grades similar to those of IT support staff. This makes it hard to recruit people with outstanding software skills and serves to further undermine the value of software development, data management, data curation, and code development. It is commonplace for universities to pay those with technical skills, such as clinicians or accountants, something closer to their realistic market salaries. Universities should develop pay scales for developers in the same way that they have done for clinical academics, recognising the specialist skills and outside options of those they are seeing to recruit.

Open 24. Create a working group to develop an attribution model for code and data

Academic behaviour is inevitably driven to a greater or lesser extent by crude metrics. There has been extensive discussion about recognising contribution to code and datasets, and attributing credit where work is re-used. None of this has delivered concrete change. Examples of options include: Direct Object Identifiers (DOIs) for code and data; recognising citation of these objects in the h-index and other commonly used metrics; supporting publication fees for publishing code and data, or links thereto, as placemarkers in conventional academic journals for citation and recognition. A Working Group should be established with support from UKRI to review prior work and set out options for an implementable set of detailed proposals within 12 months in an openly published document for consultation. This working group should consist of policymakers, senior and junior RSE experts, senior and junior researchers, representatives from academic publishers, representatives from funders, and representatives from organisations producing data for re-use such as the prospective longitudinal research cohorts. This group should produce an implementable model within 24 months and monitor implementation, in order to create and maintain incentives and recognition for those sharing access to data and code.

Open 25. Clarify the need for authorship for software developers and data scientists as equal core contributors

There is currently a wide range of norms around whether those making deep contributions to the software elements of a research project warrant authorship. Many of the current working practices are also inconsistent: “statistical programmers” are included by many groups; but not if there is a very large number of them; or if they contribute to a wide range of outputs; and so on. Current norms around authorship also tend towards regressive attribution, being more likely to include senior authors than junior contributors. Overall, having discussed this issue extensively, it is clear that there should be a presumption to include software developers and data scientists who have contributed to the delivery of the paper in authorship, not least because this work commonly entails a wide range of creative input to deliver the work informed by deep technical knowledge of the analysis, but also in very many cases the clinical domain, and the statistical context, and similar issues. Where this is argued to conflict with other current documentation on principles such as guidance from the International Committee of Medical Journal Editors - which is itself commonly breached across a range of other topics - this should be robustly addressed and discussed. Ownership of this task is complex, reflecting slow progress on this issue: initially this should be managed by the same group as above, or a prominent organisation from the RSE community.

Open 26. Proactively address sharing during the pandemic

Academic researchers have played an essential role in the response to the global COVID-19 pandemic. Yet, despite the global nature of the emergency, not all research, code, and other outputs have been made openly available for others to re-use or learn from. In some instances there may be good reasons for this, but in others it may simply be that the barriers to sharing have proved too high for some research groups - for example, paying for GitHub or for open-access publication. University administration teams and innovation offices should discuss with researchers providing research outputs on COVID-19 whether they can share code, methods, documentation, libraries, or more, for recent and future outputs.

Open 27. Academic journals should be encouraged to make code-sharing a requirements

In recent years, academic journals have begun to embrace the idea of open knowledge, introducing options for open-access publication, allowing for pre-printing and self-archiving, changing the copyright for articles, and - in some instances - introducing data sharing requirements. Less progress has been made in the domain of code sharing. This can be detrimental, as even reviewers sometimes do not have access to analytic code to check the work when making decisions about which articles to publish. Journals should begin consultations with the wider academic community about the development of code-sharing policies.

Open 28. Embrace Research Software Engineering with three Data Pioneer groups leading by example

The RSE model is strong and proven, but under-used in the field of health data. RSE groups can work collaboratively with teams to help them adopt RAP methods with hands on assistance and with training. They can help with larger code projects to identify the right level at which to "abstract" the task into re-usable functions. RSE groups developing more resource in universities will mean that they naturally develop more influence over the culture of how research is done, and rewarded. This is best achieved by centrally resourcing three RSE groups via UKRI focusing on health in three universities, prioritising teams where there is established capability to augment. This group of three Health Data RSE groups should meet regularly to coordinate advice to UKRI, the NHS, and the university sector more broadly, on how to expand software capability and productivity in the broader health data community.

Open 29. Use research funding to drive modernisation around better use of code and data

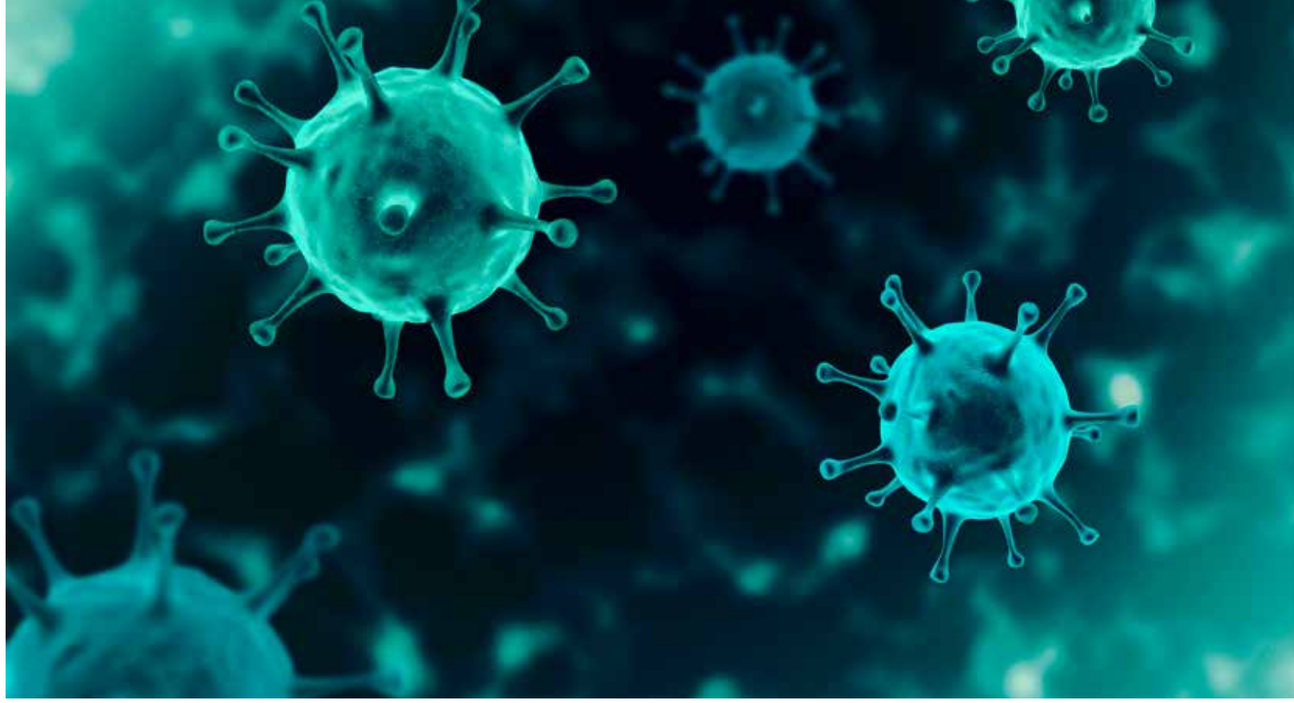
Research funders have a unique ability to shape the behaviours, priorities and delivery of the research community. The following recommendations will help them to drive positive change around adoption of RAP and modern, open computational approaches to data science in healthcare.

Open 30. Make code sharing a core feature of all standard funding contacts

Academic funders including NIHR and UKRI must facilitate and require open code as standard practice. There will be some circumstances where making this mandatory will be inappropriate, so all contractual obligations must include an exceptions process, whereby applicants can openly argue for exceptions in single cases, according to a pre-specified set of criteria, as discussed elsewhere in this Review. Exceptions should be agreed before work is funded. Applicants for funding for quantitative research projects should be required to state how they will share code the project, as some funders require applicants to do with data and dissemination.

Open 31. Provide guidance and training on RAP and code sharing

The training above covers broader issues around RAP and code sharing. Some researchers currently funded by for example NIHR and UKRI may be willing to share code but require bespoke training and support to do so. To help these individuals with this transition, funders should provide guidance and funding to attend training on 'good enough' code sharing, alongside guidance and examples of 'perfect world' code sharing. This should include supporting existing work, such as the Turing Way, RAP, and other similar projects.



Open 32. Fund a fellowship programme around Code for Health Data

There is currently a need for more teams and individuals who combine skills in computational methods, alongside domain knowledge around epidemiology, NHS analytics, and EHR data. This can be addressed by applying the normal mechanisms for capacity building, especially fellowships which bring independent status within the university sector for individuals and a field; support career progression; and support individuals to make their own choices about where they can best contribute and innovate collaboratively alongside pure research colleagues, to get away from a paradigm of developers being seen as "instructed by" researchers. UKRI and/or NIHR should fund the following:

Fellowships for software developers in health data

UKRI in particular already have some fellowships for developers and engineers; this should be energetically expanded into health, with a ring-fenced count of fellowships for this purpose. To maximise the talent pool, it is crucial that access to apply for this funding is open to all, and not limited to applicants from a specific set of academic groups or educational institutions.

Entry Fellowships scheme for developers from other sectors

To solicit inward movement from other sectors it will be valuable to have a small range of fellowship schemes were new entrants can have their salary covered for the first year of their inward movement and training for domain-specific knowledge such as epidemiology, EHR structure, NHS service analytics, and related activities.

Training fellowships in computational methods

To help individuals in conventional academic positions develop their computational skills it will be valuable to have a small range of fellowships, ideally attached to a curriculum of training, for those developing such skills. To maximise the talent pool it is crucial that access to apply for these fellowships is open to all, and not limited to applicants from a specific set of academic groups or educational institutions.

Open 33. Open funding calls for projects and programmes around Code for Health Data.

Work related to technical infrastructure, data management, and TREs is not universally regarded as legitimate academic or methodological activity. There are multiple sources of funding for Individuals and teams to address specific single clinical research questions, but almost no open competitive funding for the development of code, infrastructure, or innovative methods in this space. It is crucial that great code and data infrastructure is developed in close collaboration with the delivery of strong single academic research paper outputs. However strong code is not produced when this aspect of the work is left to fend for itself as a junior party to single topic research projects, especially during a period of transition towards more computational approaches. UKRI and/or NIHR should launch an open funding call specifically for open code projects in health data science, and consult closely with the Wellcome Data for Science group on the best means to achieve this. An initial list of example projects is given in a box at the end of the TRE section for initial guidance and illustration only. To maximise the talent pool it is crucial that access to apply for this funding is open to all, and not limited to applicants from a specific set of academic groups or educational institutions. The following is recommended for any funding programme.

Open 34. Treat "Data Infrastructure" as Open Code

Recent work from UKRI increasingly recognises that modern infrastructure is less about computers and buildings, and more about about code and teams. It is crucial that data infrastructure is handled in this way, with delivery of open code at its core, accompanied by detailed technical documentation alongside it, consistent with RAP and the computational approaches outlined above. It is equally crucial to recognise that this approach can and should viewed as modular (with different re-usable elements for different tasks produced by different teams) and methodological (with approaches to specific data management, data analysis, and privacy preservation tasks developed iteratively through innovation and delivery). Most importantly, the system must steer carefully away from procuring infrastructure such as TREs as "black box services" where only comms material and finished academic manuscripts can be seen from the outside; this is covered more in the chapter on Trusted Research Environments. Some initial outline examples of the types of activity that funders might solicit or respond to in the health data space are given in the box at the end of this section.

Open 35. Use Open Competitive Funding for code projects

It is vital to support an open collaborative ecosystem where those with the best ideas and delivery can compete to propose the best approaches to diverse challenges in data management, analysis, analytic platform components, and so on. An approach where one organisation, or a small number of organisations, is viewed as the sole supplier will not lead to excellence. Open competitive funding, where all can propose projects, is a key route to ensuring we identify and resource the best ideas and teams.

Open 36. Review prior delivery of open code by applicants

During a period of transition to new ways of working it is important that those leading by example are enabled to drive change; and prior delivery in this comparatively new space will likely be a strong predictor of future delivery. Particular caution should be used around teams with prior substantial resources for health data research projects that have not delivered, or cannot retrospectively showcase, a commensurate set of documented code.

Open 37. Ensure experts on code select and oversee code projects

Individuals with direct technical expertise in software projects, data infrastructure, RSE, RAP and related work should lead or very strongly guide all funding for projects in this space, just as conventional academics guide awards and oversight for projects delivering single academic paper outputs.

Open 38. Ensure the objectives and outputs of all investments are open

All funding for code projects, especially those around infrastructure, should be openly and publicly disclosed, with a brief description of the amount, the recipient, the expected work and timelines, and a link to the repositories where development and documentation is anticipated to reside.

Open 39. Ensure funding for code and platforms does not get diverted

When providing bespoke funding for developing open code and appropriate technical platforms it is necessary to ensure resource is not diverted to fund single research paper outputs for which there is extensive funding through other routes. This will require oversight at a number of different levels. When applications are being reviewed, funders should make sure that they have relevant technical experts on the panel reviewing the applications - for example, research software engineers and data scientists - evaluating proof of skills such as GitHub profiles and repositories in addition to CVs.

Open 40. Avoid "regressive funding models" built around short-term bursts of funding

There is a concerning tendency for code projects to be resourced with short-term urgent bursts (for the avoidance of doubt, outside of COVID-19) with funding arrangements that require successful applicants in extremis to spend a large amount of money very suddenly, over less than a year, starting with only a few weeks' notice. This strongly suggests a lack of strategic thinking in the organisations offering funding, and will tend to preferentially resource large incumbents who can rapidly absorb such costs, but who may not have the best prospects of delivering for the wider community. More specifically this last-minute funding strategy actively mitigates against new entrants to the market, field, and creative academic pool, while favouring incumbents; and disadvantages those in more junior roles, or without large existing teams, who may have the strongest ideas, newest skills, and best delivery prospects. The preferred funding approach should operate on 2-5 year cycles to build capacity and open delivery, with the conventional option of 6 months from award to commencement to facilitate inward recruitment of staff.

Open 41. Focus on sustainability for software projects: set aside a third of resource for this task

Funding should work hard to identify teams and projects with broad user bases and impact, then support them to continue their work with 2-5 year funding. This should not entail unrealistic proposals for commercialisation of data management and analysis code used solely for research and health data analytics unless there are clear grounds for a standalone commercial spinout; and it should not necessarily entail

extensive commitments to specific new features simply to justify sustaining a project. Iterative improvement and sustained delivery are sufficient goals in themselves, build capacity in the system, and build a broader ecosystem of productive outputs. A third of any budget should be set aside for sustainability.

Encourage open working through Trusted Research Environment design and implementation

These issues are also discussed in the TRE chapter.

Open 42. All Trusted Research Environments for NHS data must facilitate and require code sharing

All national TREs should be designed such that researchers are able to use services such as GitHub and Gitlab; and ideally require such that code is shared by default. This must be a core part of the core design of any TRE for NHS data, and should be a key feature of any accreditation or recognition criteria for a TRE.

Open 43. TREs themselves should be built on principles of RAP and open code

It is very reasonable for TREs to use general purpose closed commercial products (such as enterprise database tools) where these reflect the best procurement choice. However, any tooling built for the TREs themselves - especially the components for data management, analysis, and task execution - should be built using the principles of RAP and open computational working as above, with deep technical documentation and open code to facilitate access, usability, accountability on delivery, and modular development and augmentations within the environment itself.

Open 44. Produce Clear Guidance on Disclosure Risk and Open Code

If code is not written thoughtfully, and consciously, there are small risks of disclosing small amounts of personal information. Release of open code from TREs should efficiently manage disclosure risks, and TREs should give clear and enforced guidance to users on the requirement for their code to not contain any disclosive information, alongside all their other guidance on secure analysis of sensitive personal data. Writing non-disclosive code should be an absolute requirement for those requesting access to NHS patient data, with a robust and public exceptions framework for specific situations where there is no alternative, with appropriate alternative steps to manage code sharing and privacy. Those accessing NHS patient data should be required to commit to this, and demonstrate understanding of how to achieve it (for example in the training and tests required for access).

Chapter 03

Privacy and Security

Goldacre Review

Summary

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 inaccessibility 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 school gate, 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 (GPDPR) 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 (GPDPR) data collection. 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 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 extraction programme. 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.

There are 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.

Background

The system has an admirable and energetic ambition to broaden the use of health data: to get access to more health data, more conveniently, and to get a wider pool of analysts, researchers, and innovators working in it to deliver improved patient outcomes. While the benefits of working with data are widely discussed, and very well established, there are also challenges. Throughout the process of this Review the team has heard extensive concerns of two kinds, which are - only superficially - in tension with each other.

Firstly, the team heard extensively from campaigners, patients, policymakers, and professionals about the need for patients' privacy to be respected, about the need for medical records to be handled confidentially, and the substantial shortcomings in current ways of working with health data, such as disseminating detailed data out to multiple locations where subsequent activity is less clearly monitored. These concerns are legitimate, and must

never be dismissed. Managing a health service effectively, or delivering high quality research at national scale, requires that analysts have access to the most detailed information, in the GP records of every individual in the country, to do their good work; but 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. This is an extremely serious undertaking whose gravity must never be underestimated, if the NHS is to maintain trust and active enthusiasm from patients and the public.

Secondly, and at the same time, the team has heard extensive, detailed, and widespread concerns from researchers, funders, NHS analysts, and senior leaders in the health service and government, that the administrative, regulatory, practical and legal aspects of Information Governance (IG) are obstructive, duplicative, expensive, slow, inefficient, confusing, inconsistent and disproportionate to the risks presented by their work. Ultimately, many feel that the current approach to IG fails to account for the good that they can do with better access to data, and that their access should be made more straightforward.

In the past these two challenges have been presented as a dichotomy, or a trade-off, whereby greater access to data automatically means more risk to patients' privacy. In our view this is incorrect. It is entirely possible to achieve both broader, faster data access, and better protect patient privacy. This can be achieved by moving away from outdated working practices, whereby data is disseminated inefficiently and insecurely to a wide range of different sites. Instead, the NHS can ensure patients' privacy – but also deliver more efficient and higher quality analysis – by moving to a paradigm where all analytics work is done in a small number of secure analysis platforms, where the researchers come to the data, and export only the answers to their analytic questions, rather than individual patients' medical records.

"The set-up of siloed, and slightly out-of-date data was fine for 25 years ago. This is no longer a viable model. We can have multiple eyes on the data in near real-time."

- Interviewee

These platforms are generally known as Trusted Research Environments (TREs). Despite their patchy adoption in healthcare, TREs are the norm in many other settings where analysts work with disclosive data. For example, the Office of National Statistics has run a "Secure Research Service" (SRS) since 2005: this TRE is the only way that any analyst is permitted to work on data from the UK Census, because it is widely recognised that census data remains disclosive of confidential information, even when direct identifiers such as name and address removed. To understand why and how TREs can provide the NHS with a means of simultaneously providing analysts and researchers with broader access to data, and patients with better privacy protection, it is necessary to first consider and address, openly and in detail, the privacy challenge inherent in working with detailed health data; before considering how these privacy risks can be managed with the use of TREs.

It is important to note that nothing in this section should be taken to imply that there is any impending privacy catastrophe around NHS records. There is certainly a pressing need to deliver better, more efficient, more secure, and shared analysis platforms before access is granted to more granular data (such as the GP data extraction) and before access is granted to a wider number of users. But there is no need for any current systems to be urgently switched off. During this Review, one of our key recommendations – that the national GP data extraction should only be made available through a Trusted Shared Research Environment – has already become official government policy. This has been met with strong support from the professions and campaigners, and means that the system has already begun the important work of mitigating any new emergent risks, and building more efficient working practices.

This section sets out the practical nature of the privacy challenges inherent in working with NHS data, in order to help all involved make informed choices about the best mitigations, including TREs, and to help give context for the current Information Governance framework, and how it can safely be modified in a way that addresses privacy concerns while also delivering a rapid expansion in high quality analytics and research from NHS data.

Privacy risks in detailed health data

In this section as with many others the focus is on electronic health record (EHR) data, as this is the most widespread and readily usable data in the NHS. As discussed in the introduction, NHS electronic health record data can be conceived of as a series of rows, each of which contains a patient identifier, a date and time, a location, an event code, and sometimes another associated "variable" or "value". While this detail makes NHS data very powerful, it also presents 2 new challenges. Firstly, as already discussed, the sheer scale and complexity of the data makes it difficult to use efficiently. Secondly, the detail presents an important challenge around patient privacy.

The current norm when working with large NHS datasets for analytics or research is that the records are "pseudonymised" before being disseminated onward for use on the analyst's own computer: this might be some kind of Data Access Environment within their organisation, or a standalone laptop or desktop computer. Pseudonymisation means that direct identifiers such as name, NHS number, street address, and precise date of birth are deleted from each row of information about the patient, and replaced with a unique pseudo-identifier (a pseudonym). This pseudonym is typically created by a "cryptographic hash" combining the original NHS number with a "salt" (best considered as a one-way key) to make a new identifier, replacing the NHS number on all records in a manner that is hard to reverse, but can be repeated if need be. In the table below, the fictious patient "Mary Smith" has had her name and NHS number replaced with the pseudonym DLJ9821398, and her GP practice name and address is similarly converted into a pseudo-identifier for the practice. The records arising from her single cystitis consultation would now look something like the table below.

Pseudonym	Event code	Event definition	Date, Time	Location
98A2U2T9E	324431001	"Trimethoprim 200mg tablets given"	30/6/2021 10:31am	284383
98A2U2T9E	1090711000000102	"Urinary tract infectious disease (disorder)"	30/6/2021 10:31am	284383
98A2U2T9E	324431001	" Mid stream urine sent for culture and sensitivity (situation)"	30/6/2021 10:31am	284383

This pseudonymisation process does ensure that individuals are not immediately identifiable to researchers or analysts simply looking at the dataset, by accidentally seeing the name of someone they know. It does not, however, completely remove the risk of re-identification, especially when someone is actively looking for disclosive information, because the events in the records themselves can be sufficient to uniquely identify individuals. Pseudonymisation alone does not, therefore, offer sufficient privacy protection, particularly when dealing with very large, very detailed datasets like GP records.

To understand the shortcomings of pseudonymisation alone, it is helpful to introduce the concept of "threat modelling". This is the process of systematically identifying all the ways in which researchers or analysts could pose a threat to patient privacy by re-identifying individuals. This does not imply that all researchers and analysts are unworthy of our trust: indeed, they are carefully evaluated to be trustworthy by various administrative systems. However, systems and services that rely on trust alone are not infallible, and struggle to scale. Services should be resilient to "bad actors" wherever this is practical, and resilient services are especially important in the context of plans to rapidly expand the number of people accessing NHS data for research and innovation. More than this, it is important to understand the challenges and risks around re-identification, in order to develop proportionate mechanisms – both technical and governance - for mitigating those risks.

An example of re-identification with socially accessible information

It is easiest to consider the problem of re-identification with concrete examples. 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. If one created a table to show how many people have each combination of birth month, region at various timepoints, and childbirth month, then in that table there would be many "cells" with only one individual in them, meaning that these individuals could be uniquely identified in the GP dataset with only that information. This mechanism has indeed been used to help identify individuals in health data when a large dataset was unwisely released to the public in Australia (see the paper here) demonstrating the extent to which some have previously over-estimated the protective benefits of pseudonymisation.

The re-identification of individual women through this kind of information is particularly noteworthy, as a privacy challenge, because it means an untrustworthy actor with unconstrained access to data could illegally view their targets full medical records in a pseudonymised dataset knowing only trivial information about someone's children, age, and region of residence. This is the kind of information which in many cases will be available to acquaintances, colleagues, or even strangers. Having identified the pseudonymous identifier for one individual, any untrustworthy user of the data could then see all the other entries in their target's medical history, which are also recorded in the same dataset, including all the things they didn't already know.

An example of re-identification with information about someone's medical history

It is common for people to disclose information about their medical problems or treatments, or have it disclosed about them. A comparatively small amount of information about one individual's medical history can often be enough to create a unique match in a large pseudonymised health dataset. For many people in the public eye, this kind of information may also be in the public domain. It is commonly reported in the media that a person in the public eye has had a particular kind of medical procedure, at a particular time, in a particular location. The precise medical procedure itself need not be particularly unique for this to be sufficient to identify that individual: the combination of the procedure, the approximate dates, their approximate location, and their approximate age is likely enough to create a unique combination, even if one only considered the week in which each event occurred. Having identified the pseudonym for one individual, any untrustworthy user with unconstrained access to the remaining data could then, once again, go on to view everything else in that individual's medical record: they would be breaking the law, but depending on the setting, they may not be detected, or prevented from doing so.

The consequences of re-identification in NHS data

It is important to be open about the gravity of the risks with NHS data, in order to demonstrate to the public that they are being taken seriously, and to develop mitigation strategies that are both credible and effective. Health records are some of the most sensitive information that is stored about an individual. Some individuals may be so relaxed about the contents that they would be happy to see their entire record in the public domain. Many patients have entries in their medical record that they would strongly prefer were not public knowledge, or known to someone without their permission, whether that is a partner, an ex-partner, a colleague bullying them at work, a journalist, a stalker, or a private investigator. This could include issues they might find embarrassing, such as incontinence, or a mental health problem; it could include sexually transmitted diseases after the commencement of a monogamous relationship; it could include region of current residence for someone seeking to avoid an ex-partner. The leaking of confidential medical information could be, for many individuals, a life-changing and catastrophic invasion of their privacy.

Anxiety about the disclosive nature of individual medical records has driven many of the objections to large scale data collections - such as the recently planned GP data for Planning and Research data collection (GPDPR) and the earlier care.data project - from professionals, campaigners, patients and commentators. Patients have a legitimate desire to see their medical records kept confidentially, and their privacy protected; the challenge is to achieve this, and prove trustworthiness, while also ensuring that data is accessible for vital work that can improve everyone's health and wellbeing.

The additional challenge of large aggregated datasets

Importantly, the risk of re-identification in pseudonymised data also increases, as the dataset grows to cover a larger proportion of the total population. The increase in risk from larger datasets is driven by 2 important factors. Firstly, a large dataset is inherently a much more attractive target: it is almost guaranteed to contain information about the target of interest, and therefore it is a security asset of greater value for misuse, illegal access, but also cybersecurity attack (which is itself another reason to minimise the number of locations storing large datasets of pseudonymised but re-identifiable data).

Secondly, the maths of disclosure and uniqueness mean that individuals can be more confidently identified in a dataset that covers a larger proportion of the total population of interest. For example, if an untrustworthy user, with 3 characteristics about their re-identification target, found a unique match in an NHS dataset covering a 5% sample of the population, then (assuming no additional knowledge) they could only be 5% certain that the unique match really is the person they are looking for: because there might be 19 other matches in the remaining 95% of the population that they are unable to search across. However, using data on 100% of the population, if a unique match is found, then the untrustworthy user can be 100% certain they have found the target of interest.

For clarity, this does not mean that a small population sample of records is inherently safe for dissemination, or other less stringent security handling: because the untrustworthy user might also have other reasons to already know a priori that the combination of features they know about their target is so rare as to be unique, meaning that they can still confidently identify their target, if that individual is present in the sub-sample dataset to which they have access.

The benefits of pseudonymisation

None of the above should be taken to mean that pseudonymisation is without value. It remains an important technique, because removing the names and addresses does meaningfully prevent one very narrow form of re-identification: the accidental viewing, in context, of a record that belongs to someone known to the analyst. For example, if an analyst is creating some summary information tables to describe how effectively clinics in their region manage early pregnancy, they will be using information from the records of women in their region who have had a recent positive pregnancy test recorded with local services. If the data has not been pseudonymised, and they see their colleague's record flash past, that record might catch their eye. If their colleague has not yet told them that she was pregnant earlier this year, then the analyst has seen something they cannot un-see. Preventing this scenario is useful: but that is the only privacy threat in detailed health data that pseudonymisation alone can meaningfully address.

The likelihood of re-identification in NHS data

Almost all security initiatives entail a degree of cost and inconvenience: adding an extra lock to a front door will make it more secure, but it will also make it somewhat slower to enter and leave the home. Because of this, it is common and reasonable to hear objections from those who use data to new proposals for improved security - whether those are administrative, governance, or technical proposals - as those may bring some inconvenience for people delivering high value work.

During this Review multiple stakeholders from the academic research community expressed the view that the privacy risks described above are theoretical, or that attacks on data are vanishingly rare. It would be helpful if this were true, but it is already well established that large non-health datasets are commonly misused, both through malice and curiosity.

There are numerous examples of misuse of smaller local GP and NHS datasets, including GP practice staff illegally viewing the medical records of ex-partners, or women they went to school with. During this review the team were also confidentially given examples of celebrities' and colleagues' medical records being accessed illegally by staff in NHS environments. Typically, these transgressions have been met with low-level penalties.

Overall, it is proportionate and appropriate to seek to implement technical and governance processes to ensure that pseudonymised patient data is handled securely.

Similarly, celebrities and others often have information about them leaked from national non-health databases to the media. Some have argued that there are no specific cases where a medical researcher has been identified as misusing health data. This may be the case -it may not be – but the current pool of researchers with access is historically small, and in many cases the storage of these records (in multiple locations, including analysts' own laptops) is not subject to the same efficient audit processes as police records, or GP records in one practice.

Overall, it is proportionate and appropriate to seek to implement technical and governance processes to ensure that pseudonymised patient data is handled securely, for a number of reasons: re-identification and leak of disclosive information is possible; it would have very bad consequences; evidence from various settings shows that misuse does happen, even where there is a reasonable chance of detection; and it is practical to manage data securely while also granting access straightforwardly.

This does not mean that such misuse is a foregone conclusion, and it certainly does not mean that data collections such as the 'GP data for planning and research' dataset should be cancelled, because they will bring spectacular health benefits for patients. Rather, it simply means that such misuse must be recognised as a genuine risk, and managed. As a general principle, all pseudonymised national detailed health datasets should be treated as if they have name and address in the clear, both practically and in our governance, regulatory, and legislative frameworks.

Other forms of risk mitigation

As discussed in the next chapter, Trusted Research Environments are the best mechanism to address the re-identification risks inherent in pseudonymised data, because of the privacy and productivity benefits that a good TRE can bring. However, it is useful at this stage to consider the other mitigations sometimes considered or implemented to protect patients' privacy that fall short of a robust TRE. Many of these have value in context, including when preparing data for use in a weaker form of TRE, as discussed in the next chapter. However, most are not a panacea on their own.

Removal of "sensitive codes"

This method entails the deletion from the record, prior to onward dissemination, of certain specific groups of codes regarded as particularly sensitive, intrusive, or disclosive. These typically include codes on sexually transmitted diseases, terminations of pregnancy, and mental health problems. While superficially appealing, this approach to preserving privacy creates a range of subsequent problems. Firstly, there is little universal agreement on what information is or is

not stigmatising: for example, some people might feel that urinary or faecal incontinence should be regarded as embarrassing information, but this falls outside of typical "sensitive code" lists; and many people would simply prefer that everything in their record was managed securely. Secondly, this approach obstructs important analyses that use whatever codes are deemed sensitive, and thereby inflicts unintended harm on whole categories of patient by depriving them of the benefits that come from important research or service improvement work. Were mental health codes to be withheld, for example - as is often suggested - then this would prevent us better understanding the causes and treatments for these conditions, whether they increase patients' risk of developing other physical health problems, whether there is variation between organisations or regions in how patients' mental health problems are managed, and so on. Lastly, a remote and secure analysis of someone's health data can be the least intrusive means to do research or analysis on the medical events that some regard as sensitive, and therefore may be particularly warranted for such issues: for example, if conducting an analysis on whether termination of pregnancy is associated with later health problems, then a telephone call, letter, or knock at the door from a researcher with a questionnaire survey and lengthy consent form may be substantially more intrusive and undesirable for some people than their health records simply being included among millions of others for a robust national analysis of data. Overall, the removal of sensitive codes is something that should be considered in context alongside other mitigations, for specific types of analysis where it can be shown to impose no active harm on patients, but not regarded as a foreground technique for mitigating privacy risks.

Data minimisation

A key principle in the General Data Protection Regulation (GDPR), this method entails reducing the amount of information shared about each individual patient, down to the bare minimum necessary to conduct the desired analysis. It can be done in a variety of ways. Sometimes it is done by releasing only specific types of code. For example, in principle, at first glance, a researcher working on a project to describe the natural history of rheumatoid arthritis needs only to see the codes pertaining to the rheumatoid arthritis (RA) codes in each patient's record: the codes describing the dates and details of each RA treatment, referral, blood test, examination, operation, and so on.

However, in reality, problems often arise. For example, the analyst in this project needs to see the outcomes that can arise for patients with RA: these won't just be direct RA codes, and they won't just be joint related. A wide range of medical problems can arise from RA, and other non-RA problems can be confused with RA. Furthermore, patients' medical problems often interact, and make each other worse: so the analyst will need to know about those too. For most types of research looking at risk factors for a given condition (a very common variety of analysis) analysts will want to know about many aspects of each patient's medical history, in order to "adjust" for them in the statistical analysis. Overall, the approach of limiting the types of field that are accessible can be helpful, but again it is often not a panacea.

Another approach is to compress the record, reducing the granularity of all the information, and sharing only the information that the analyst actually needs to do their work. This can be done in a number of ways, using similar techniques to those discussed in the chapter on Data Curation. For example, a researcher looking into life expectancy among patients who have had an abnormal heart rhythm may not necessarily need to know the precise dates of every separate relevant treatment, diagnosis and referral for every patient: they may be able to deliver their analysis to the same quality knowing only, for each patient, overall, whether they have or have not had an abnormal heart rhythm of a given type in a given year. This information is derived from the full data, using a set of complex rules, but is much less disclosive. If an analyst was given only this broad brush information about each patient, then it would be substantially harder for them to identify an individual using the methods described previously, and so the dataset would be substantially lower risk for dissemination.

This approach has merit, in the context of a wider ecosystem including TREs; it is also a crucial element of how to best transfer information between TREs where this may be necessary. However, it brings many practical challenges, and in some respects it only shifts the problem along the road, because the task of minimisation is effectively the same as data preparation. As discussed in this chapter, and extensively in the chapter on Data Curation, data preparation is widely regarded as the bulk of the work in any analytics project. Furthermore, it is a highly skilled technical job that can be done well or badly, and there is often substantial reasonable disagreement on the best way to do it. As testament to this, throughout the review the team received many complaints from users that when they requested datasets to be prepared for them by data controllers such as NHS Digital they were only able to communicate those requirements and data transformations verbally, or in the form of written descriptions, rather than in unambiguous "code": this leaves room for ambiguity that worried researchers, and resulted in data management work that is crucial to their analysis being done with unknown methods, outside of their oversight and control. Data minimisation is therefore only useful insofar as it genuinely makes data less disclosive; and is only feasible insofar as the system is able to adopt more contemporary approaches to dataset requests in code, as discussed in the chapter on Data Curation. It is crucial that NHS Digital, in particular, as a key relay in brokering access to data across the system, should be able to accept technical descriptions of dataset requests in "code rather than conversation".

Even when this is achieved, as discussed in the section on TREs, there has been very little applied privacy research on what level of data minimisation is necessary to protect patients' confidential records when sharing access to

information about them: this is a shortcoming that can be readily resolved by UKRI/NIHR embracing a technical and "methodology and code" approach to improving TRE provision with the academic community.

Sub-samples of the population

As discussed above, a smaller sample of the population represents a somewhat smaller risk of re-identification. As many analyses on common medical problems can be practically delivered with only a small random percentage of the total population, it is reasonable to ask that smaller samples are used. However, this approach still has shortcomings. Firstly, it is often necessary to have the greater precision that comes from using the largest possible population. Secondly, a local analyst looking at local data is likely to need the full local population. Thirdly, the privacy benefits of sub-sampling are only relative. Fourthly, sub-samples currently delivered are typically "convenience samples", rather than a

true random sample: for example, the MHRA's Clincal Practice Research Datalink (CPRD) system for GP data extraction and dissemination takes all its records from a known list of individual GP practices around the country. Despite these shortcomings, true random sub-sampling can be helpful in context alongside other safeguards, for example in a weaker "remote desktop" form of TRE that cannot so easily restrict or detect misuse of data, or share informative public logs.

Data perturbation and "synthetic data"

This is a very large field of activity, but in brief the approach entails adding random noise to the dates, variable values, or other aspects of each patient's record, in order to reduce the disclosiveness of the data, making it harder to identify an individual in it. This has been an area of energetic research, as it lends itself to mathematical modelling work that is interesting to conduct and publish. However, it presents very substantial challenges, as the overall objective is to introduce just enough noise to make individuals no longer identifiable, but preserve the true data sufficiently that the research findings derived from it remain correct. It rapidly becomes very complex in implementation, as the perturbation algorithms often hit real world problems specific to a given medical domain: for example, there may be a need to preserve the date delays between specific events, otherwise the data becomes hopelessly unrealistic and unusable; but that may conflict with a broader requirement to perturb dates; and requires an informed user to spot that it is an important issue to address; and so on. It also poses practical, methodological, and security engineering challenges outside of the single perturbed dataset: for example, if multiple copies of the same (or similar, or related) dataset are released, each with different random noise added to the dates and values, then the true values for each patient could plausibly be interpolated. Overall, this kind of "synthetic data" is problematic for real analyses, but can be useful for other tasks, for example as a resource on which analysts can develop analytic code for subsequent execution against real data; or to examine for training purposes; or as a service to help new arrivals in a field evaluate the feasibility of using a given dataset for a given purpose.

Emergent methods

There are various new methods emerging: dominant among these is "homomorphic encryption" – a technique that theoretically enables analysts to analyse encrypting data without decrypting it. During the course of this review some non-technical staff expressed substantial enthusiasm about the possibilities of this approach; however the team has found no credible implementation of homomorphic encryption to manage analysis of electronic health records at scale in a way that would meet the needs of service and research users while also addressing key outstanding patient privacy challenges; and as with many theoretical approaches to privacy, a practical implementation of a given theory can often entail compromises on the privacy preserving elements of the underlying principles. In general, new theoretical models such as this can hold out the inspiring and relieving prospect of a new single piece of commodity technology that can solve complex organisational and technical problems. Their benefits and costs should always be evaluated practically and with an open mind: overall homomorphic encryption is currently best viewed as a valuable and interesting blue skies research and development project, that should be separately accounted from any endeavour around delivering practical services for NHS data.

Contracts and evaluation of users

Trust and contracts have historically been a key safeguard used by the NHS and the broader system to protect patient privacy. Essentially, each user requesting a substantial download of potentially re-identifiable patient data is evaluated to determine whether they and their host organisation are able to manage the data, trustworthy, and able to commit to not misuse the data. These evaluation processes often include multiple organisations, multiple committees, and long delays, as discussed in the chapter on Information Governance. Sometimes there may also be audits of how the data is stored and used, but with variable detail and aspiration. As with pseudonymisation, this approach has substantial value, and cannot be dispensed with; but as with pseudonymisation, it cannot be relied upon exclusively. It can also be extremely slow and frustrating to navigate as a data user: the team received multiple complaints of processes taking years to complete, being opaque, and appearing to end-users to be arbitrary in places. The principal security shortcoming of this method is that it relies on assumed trust: when large volumes of data are transferred, it moves out of the direct control and oversight of the NHS, and it becomes substantially harder - if not impossible - to confidently and comprehensively track what is done with the data, or ensure that it is not misused. Here, misuse entails 2 distinct categories: gross misuse, such as re-identification of individuals; and the likely more commonplace phenomenon of people running analyses on NHS patients' data for which they do not have permission.

Analyses without permission

During the Review the team was told repeatedly that "quickly" running an analysis outside of the user's current permissions is commonplace, when they have a downloaded copy of a large volume of NHS patient data: it is likely that this is the most common form of data misuse, and likely a lower risk category than many others. Two main categories of unpermitted analysis were alluded to: each is a reaction to restrictions in the current technical and administrative frameworks; each presents different risks.

The first category was analyses conducted outside of the user's permissions, but on data the user already held: for example, a researcher might conduct an analysis on detailed GP data to evaluate the feasibility of a new research project, or to see the likely answer from a full analysis, prior to going through the workload of a formal application to conduct the project. While this is low risk for disclosure, it does illustrate the lack of controls under the current model of "data dissemination"; it may result in "publication bias" whereby salient negative findings are less likely to be published; it may compromise income to the NHS or other data services operating a "fee per project" model; and it may reflect the risk of unethical research being conducted without permission "under the radar".

The second category was analyses conducted outside of the approved platforms. More than once the phrase "everyone has a secret laptop" was used to denote that sometimes people would move a volume of person-level data onto a machine where it could be managed and analysed more effectively than on the officially supported environments and machines. Typically, this was justified on the basis that their local IT team imposed constraints that the analyst regarded as unreasonable or irrational, such as preventing the use of a standard modern data science tool that they viewed as secure. The fact that it is readily achievable to move large amounts of individual level data from one machine to another without detection demonstrates the extent to which data movements cannot always be readily controlled once they leave a TRE. The fact that many analysts felt it was reasonable to work in this way – and happy to disclose it in discussion - reflects the security problems caused by poor technology outside of Trusted Research Environments; and the extent to which individuals may sometimes lose respect for security guidance when they view it as impractical or irrational.

The need for better research into privacy preserving techniques

This section is principally focused on providing background on the risks, and mitigations, that TREs and IG aim to address; relevant recommendations are in those sections. However, this overview does shed light on one important issue. There is very little applied, practical research on the extent to which different mitigation techniques used above – in particular data minimisation – can reduce uniqueness and re-identifiability in NHS records. This means that teams managing data access requests in organisations such as NHS Digital and elsewhere are commonly applying rules of thumb, or intuition, to inform their proportionate responses. It would be better for the whole community – patients, researchers, and those serving their needs – if this were addressed with a modest programme of applied health research from a national funder.

Summary

Overall, there is no doubt that the vast overwhelming majority of researchers and NHS analysts are trustworthy. However, there are genuine risks that must be acknowledged, and mitigated, in an open and credible way to build trust from patients and the public.

The historic approach to risk mitigation has relied on approaches built around "trust and contracts" that cannot scale, and barely meet user demands. The system itself is extremely slow, not just causing frustration, but also blocking valuable research and analysis projects outright, or delaying them for years until they are abandoned. The individuals administering the system are, from our extensive discussions, often very anxious about each individual data release or download that they approve. There is no prospect of this system being somehow liberalised overnight, and there is a clear urgent need for TREs to deliver a safe and efficient solution to NHS analytics at scale for two reasons.

Firstly, the datasets themselves are now larger, more disclosive, and more vulnerable to re-identification, than any previous resources: the proposed 'GP data for planning and research" dataset covers the entire medical history of almost every individual in the country, in very great detail, with many personal and non-medical issues implicit in its contents too. Datasets on this scale, with unprecedented depth and coverage, present exciting new opportunities that mean they should be strongly supported; but they also present new risks that require new mitigation.

Secondly, the pool of researchers and analysts is now larger than ever before and, for good reasons, should continue to grow. The tools used to analyse data have evolved over time, and become very widely accessible. In the past, these datasets were extremely hard to move and analyse: only a very small number of people had access, and almost all had invested a huge amount of their professional lives in getting to the point where they were able to work on NHS data, each with a huge amount to lose in the case of any misuse. Today, an unremarkable home computer can have the power to store and analyse vast amounts of patient data. We are all living through an exciting explosion in data science as a profession globally: many of these gifted individuals will move from one sector to another, applying the same data science skills at different times in health, civil engineering, transport, and so on.

Governments are alive to the benefits of better use of data, with this reflected in multiple policy documents in England, the UK, and globally. This explosion of tools and workforce can bring spectacular benefits to patients from ever larger numbers of innovators and analysts from the private and public sector. There is a widespread, admirable and vital ambition to widen the volumes of data, and the number of people able to work on it, driven by a recognition that more people will deliver more life-saving insights. It is inconceivable that trust and contracts alone can scale to manage such a huge expansion in

the workforce with access to the most private information about 58 million people. Trusted Research Environments (TREs) are the only realistic way to safely deliver the huge expansion in work on NHS data that is already happening, and that must grow even more in time.

Recommendations

There is no new emergency, but TREs should be used, and data dissemination should not expand

The system should adopt TREs rapidly as the default approach for NHS data analysis, as per the following chapter. Recent very strong progress in this direction should be welcomed and accelerated. The largest risk comes from expanding the population coverage and granularity of data aggregation; and from expanding the pool of data users to drive innovation. Despite the public and professional concern raised about the GP Data for Planning and Research Data Collection, there should be no panic or repeat of what happened in 2013 following the suspension of Care.Data, when several unrelated data flows from NHS Digital to research users (such as HES) were suspended or delayed with no alternative plan for access. Current data dissemination work should be retired as TRE work replaces it; but there is no new or sudden privacy emergency.

UKRI/NIHR should resource applied methods research into privacy preservation

There is a deep ambition for wider access to NHS patient records. This requires development and evaluation of robust methods for protecting patients privacy at scale: a range of example areas to pursue are given in the chapter on TREs.

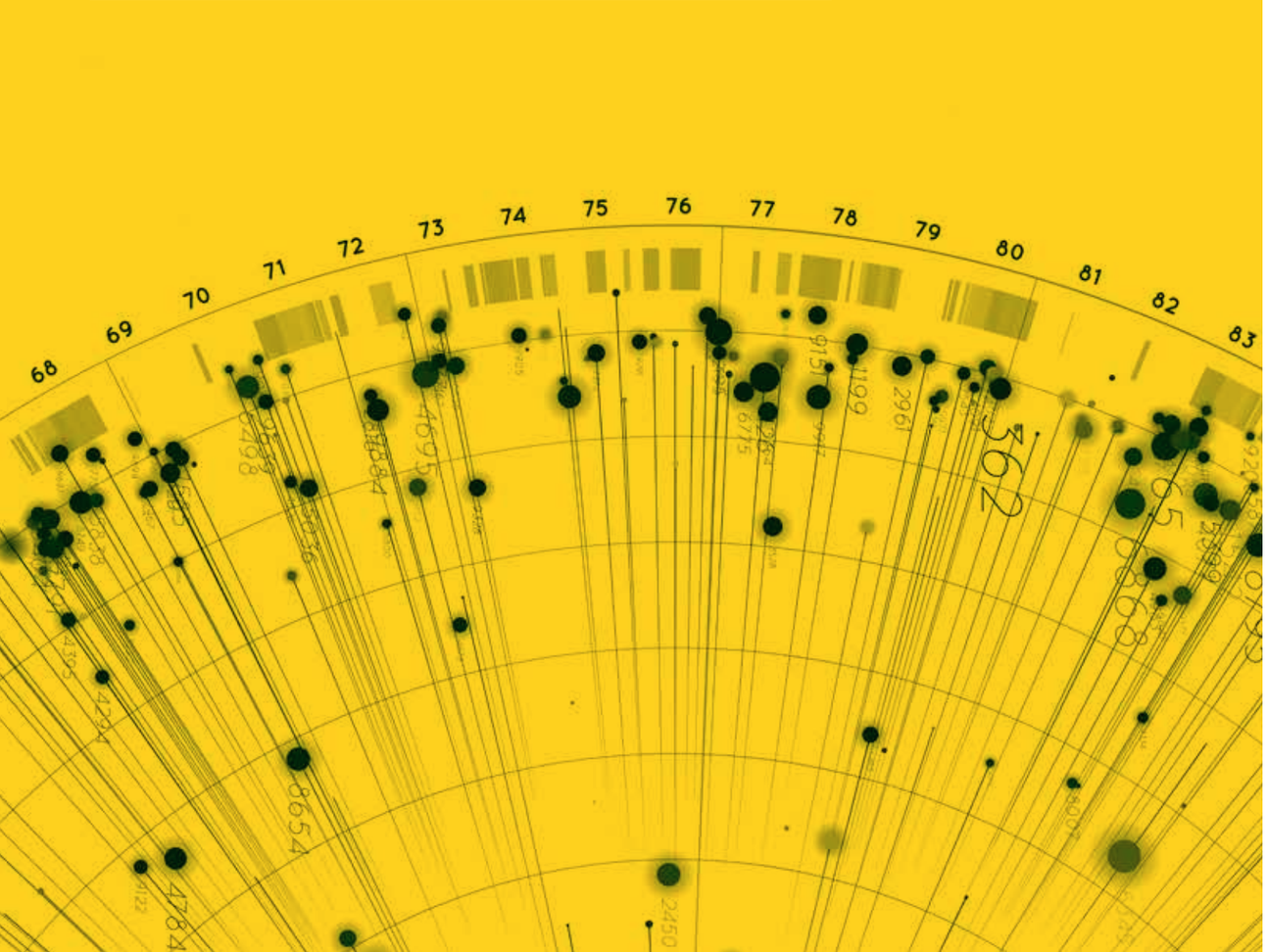
Revise the definitions of "anonymous" "identifiable" and "linked" data; add a new category of "pseudonymised but re-identifiable"

It is crucial that the system recognises and describes this category of data as a central privacy risk to be mitigated: recognising this will allow the system to make informed choices and earn the trust of campaigners, professionals, and the public. More detailed recommendations are given in the chapter on Information Governance.

Chapter 04

Trusted Research Environments

Goldacre Review



Summary

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 planned national GP Data for Planning and Research data collection 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.

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.

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. 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.

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 following 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.

Background

What is a Trusted Research Environment?

During the course of this review the planned extraction and dissemination of all NHS GP records (GP Data for Planning and Research data collection) via NHS Digital was paused, pending development of a Trusted Research Environment, with a commitment that this dataset would only be accessible in a TRE. This is a sound and proportionate move which will address a range of issues including privacy and usability of data. It has also generated a burst of activity around TREs, which have not previously received substantial attention.

In outline, a Trusted Research Environment (TRE) is a secure environment that researchers enter in order to work on the data remotely, rather than downloading it onto their own local machine. Users can extract and download the answers from their analyses - such as results tables, or graphs - but individual patients' data always stays within the secure environment. This brings many clear advantages over data dissemination. TREs are very diverse in aspiration and design, but a robust TRE can help analysts to work effectively with data while also preventing and detecting misuse of data, and providing fully open and detailed public logs of all actions on patients' records. TREs permit gatekeepers to be proportionately more permissive with access to patient data, by managing the risks around access (both in terms of privacy risk, and non-permitted uses) more effectively than dissemination.

In addition to these privacy benefits, good TREs can also provide a more efficient and collaborative computational environment for all data users, and an opportunity to make modern open working methods the simple default. Once again: it has been estimated that 80% of the work for data science with NHS records is spent on data preparation; this is currently delivered in a diverse, duplicative and ad hoc fashion, falling short of minimal RAP guidance, with different teams and individuals all using different methods and tools for even basic data management work, in different ways, in different organisations and settings, to do the same or very similar tasks, often on the same national NHS datasets such as GP data or HES/SUS. If this work is done in national or broadly standardised analytic environments, then code and working methods become portable, open, discussable, reviewable, and re-usable.

In short, TREs represent an unprecedented opportunity to modernise the data management and analysis work done across the system, allowing us to achieve the following benefits:

- Replace hundreds of dispersed analytic siloes, data centres and working practices with a small number of broadly standardised environments that facilitate the use of modern, efficient approaches to data science.
- Reduce the number of data centres, and thereby also reduce the number of cost centres.
- Reduce the number of attack surfaces for cybersecurity risk.
- Overcome local IT constraints that prevent analysts installing specific types of contemporary data science software by enabling analysts to conduct their analyses at a central online location rather than on multiple local bespoke machines.
- Create technical working environments where a smaller number of expert software developers can assist all colleagues nationally, using modern industry standard data science tools, packaging up the code for recurring tasks into adequately documented "functions" and "libraries" for easy re-use.
- Facilitate the collaborative development of highly effective interactive data tools for less skilled users with Graphic User Interfaces for safe and effective use of Point and Click tools (rather than these being an inappropriate default), using commercial and open data visualisation tools as appropriate.
- Allow (and indeed require) all data curation code to be shared with all subsequent users for review, validation, re-use, and iterative modification.
- Make modern, open, collaborative, computational approaches to data analysis the norm, facilitating Reproducible Analytic Pathways rather than duplicative, diverse and inefficient approaches to data management.

Delivering this outcome is no small challenge, but there are excellent existing templates and models to build upon. The rest of this chapter considers: the features of adequate and good TREs; the need for TREs to be shared resources; the need for a variety of TREs to meet diverse user needs; how the current explosion of services calling themselves "TREs" can be contained; and best principles around resourcing and overseeing this kind of data infrastructure.

What NHS Trusted Research Environments should look like

A TRE is superficially a simple concept: however, this simplicity belies substantial disagreement on what actually constitutes a TRE. In discussions it was clear that some users regarded a service as a "TRE" even if it was little more than a simple shared folder within an organisation's IT system, subject to controls on who could access the contents. In broad terms, before expanding into the details, it is useful to draw an early distinction between two distinct concepts. A "Data Access Environment" is an internal service where staff at an organisation can have some shared resources for secure storage and analysis of data, such as shared databases, or shared provision of computer power. A "Trusted Research Environment" is different: it has the same underlying features of shared access to data, but also a range of additional elements including a formal and disclosed process for access; a formal and disclosed process for release of completed outputs in the form of tables and graphs; a formal and disclosed process for logging activity; an audit process for usage; and so on. These are typically described as the "service wrapper" for a TRE. Overall, the single most important feature that distinguishes a TRE from a Data Access Environment is that it is shared.

"It worries me when people use the term "TRE" to describe a data access environment that is only for use within their organisation, because the need for the strict controls within the service wrapper may, naturally, feel less necessary when only colleagues are using the data. [These internal platforms] aren't genuine TREs, in my view, because they only benefit a small pool of users and do not significantly reduce the need for, or cost of, data transfers. As these environments are run independently, there is also a risk that data deposited with multiple organisations may be cleaned or curated differently in each, leading to inconsistent or conflicting analyses using the same source data. Finally, as the researchers work within the organisation maintaining the environment, there is not always sufficiently robust scrutiny of outputs, to ensure there is no accidental disclosure of personal information. I believe that a TRE should: be easily accessible to all researchers who meet (published) criteria to use the data held; ensure consistency of data between research projects; be able to monitor all analysis undertaken, to assure Data Controllers and the public that their data are being used appropriately; and provide a service to independently review research outputs, to ensure data subject confidentiality."

- Pete Stokes, Director, Integrated Data Programme, ONS

During this review the team has rapidly reviewed a range of documents from diverse organisations asserting the important design features of a TRE. Each has sound suggestions, but none are complete. Some focus more on the technical aspects, others focus on the governance and service wrapper. Alongside these diverse documents there are also many different models of TRE in existence, each with varying usability, transparency, auditability, and trustworthiness. From interviews it is clear that platforms differ substantially in their governance and service wrapper, with different approaches to the rules on permissions for people and projects, different approaches to implementing an "output checking service" to review exported tables and graphs for disclosive content, different approaches to logs or audit, and so on. Similarly, the technical implementations and design choices vary widely; however for most TREs there was little or no technical documentation beyond public relations material and PowerPoint slides. This diversity and lack of clarity indicates several features of current work around TREs: the lack of a consistent approach; the duplication of technical and governance effort; and the lack of an open commons of knowledge, meaning that current diversity of approaches is unlikely to be informative.

This diversity is compounded by the reality - often not articulated - that there are many different types of TRE user, and therefore user needs, around the system: some users require only simple tools; while other users need flexibility, and can embrace the additional technical challenges that come with flexibility.

When developing a new technical service, it is helpful to work forward from "user stories". In the case of TREs, these user stories should go beyond the direct users of the platforms themselves. A broad initial list of example user stories is given in the box below. This is an incomplete list, informed by all our discussions throughout the review, as an example of the user needs that must be met by TREs containing NHS data. Recommendations around further development of TRE user stories and capabilities are made below.

A broad list of TRE user stories including patients, professionals, and others

I am a patient / expert member of the public, and I want to:

- See what uses of NHS data have been approved
- See what analyses have been executed against the data
- See how privacy is preserved through platform design
- See how the TRE datasets provided have been minimised
- … so that I can reassure myself and others that privacy is protected, and that all purposes are legitimate

I am a funder or manager in a medical research charity and I want to:

- Signpost researchers to accessible data sources
- Signpost researchers to the code resources that make this data useful
- Improve the quality of scientific research
- Encourage reproducibility in scientific research
- Check the progress of the research I am funding
- … so that I can maximise the value of our investments.

I am an NHS provider, practice, sustainability and transformation partnership (STP) and I want to:

- See how performance metrics evaluating my clinical service were created from raw data, so that I can evaluate their usefulness
- Learn how data is used to create performance metrics, so that I can contribute

I am a data controller and I want to:

- Show all individuals contributing to my dataset that their data is only being shared for public good, securely and proportionately, so that I can build participants' trust
- Get wider use of my data, so that I can maximise the impact of our work
- Receive credit for my dataset's contribution, so that we can justify future investment for sustainability

I am a healthcare professional and I want to:

- See how my NHS patients' data is being used, so that I can learn, contribute, be reassured, and share those reassurances
- Celebrate and share how my NHS patients' data is being used, so that I can build public trust in data usage

I am the data controller for a disease registry and I want to:

- See others to use for research
- Help others to use it for research with my domain expertise.
- … so that I can maximise the value of our work.

I am a privacy campaigner and I want to:

- See the current use of data, so that I can evaluate whether it is safe and proportionate
- Identify and actively celebrate secure use of data for patient benefit, so that I can light the way for projects which don't meet my expectations.

I am an analyst and I want to:

- Access data for analysis
- Export results easily for publication and sharing with colleagues
- See other peoples' code for data management and analysis
- Re-use other people's code for data management and analysis
- Learn from other people's code for data management and analysis
- Improve other people's code for data management and analysis
- Re-run code I last used at 2 years ago against current data
- Share my code for re-use and get credit/ satisfaction
- … so that my community can deliver great insights from data

I am a software developer and I want to:

- Create great infrastructure for others, using familiar modern data science tools rather than inefficient manual processes
- Work alongside great analysts, see their work, and help them turn their single scripts into re-usable modules of code for others to use
- …. so that I can feel I am helping improve patients' health

I am an innovator and I want to:

- See what researchers and analysts are doing with data, so that I can develop tools or services to help them achieve their goals

From these user stories various features for a TRE are clearly necessary. The following list is suggested as covering most major requirements in outline, with recommendations below for further formal development.

A TRE for NHS patient data must achieve the following objectives

1. Preserve patients' privacy:

- Ensure all intentional identifiers (name, date of birth, etc) are removed at source, but recognise that this data is nonetheless re-identifiable and manage it as such
- Obstruct attempts to re-identify patients in data
- Detect attempts to re-identify patients in data
- Obstruct attempts to view disclosive information about single individuals
- Detect attempts to export disclosive information about single individuals
- Support privacy enhancing techniques, such as code development on dummy data, to minimise access to real patient data
- Prevent bulk export of identifiable or re-identifiable data
- Provide tools, personnel, training and workflow for automated and manual checking of all exported outputs to ensure they are safe and non-disclosive
- Regularly re-evaluate and compare all currently performant or realistic mechanisms to achieve the above, and ensure that only the safest are used
- Ensure all outputs are checked for potentially disclosive material by a mixture of appropriately validated automated methods, and manual checking

2. Support RAP and modern, efficient, high quality, reproducible data analysis:

- Support an appropriate range of tools for data management, analysis, and visualisation
- Permit the execution of analytic code
- Support, and ideally require, sharing and easy discovery of all code for data management and analysis
- Support, and ideally require, sharing and easy discovery of "good enough" technical documentation alongside users' shared code, consistent with minimum RAP standards
- Support the use of git, GitLab, GitHub or related tools for code management, version control and best practice in software development
- Support flexible standardisation of re-usable code for common data management and analysis tasks, where this meets users' needs
- Provide robust open technical documentation of all platform features
- Meet the needs of technically skilled users and those with fewer computational skills, providing relevant security mitigations around the latter

3. Provide a secure computing environment

- Meet and ideally exceed all relevant standards for a secure data centre containing highly sensitive, disclosive, re-identifiable patient data
- Ensure all installed tools for data management, analysis and visualisation meet security specifications, to the degree that is necessary for the security context in which they are being used
- Ensure that only users and projects with appropriate permissions are able to execute code on the platform

4. Provide a performant computing environment

- Support the rapid and scalable provisioning of appropriate resources (processor power, memory, storage, and so on)

5. Earn patient trust

- Publish the governance arrangements (including transparency notice, DPIA and relevant Terms of References of governance groups); including how decisions about access are made, according to which criteria, and who is responsible for making these decisions
- Openly disclose all code and technical methods used to preserve patients' privacy
- Publish appropriate information about all users, analyses, and their associated permissions, in near-real-time, using metadata from actual usage of the platform
- Keep, and ideally publish, detailed informative technical logs of all activity in the platform, attached to users, analyses, and their associated permissions
- Ensure all outputs of all analyses executed in the platform are shared openly, other than for pre-specified and pre-arranged exceptions
- Retain copies of all analysis results for audit

6. Be surrounded by good governance, and support this with relevant technical features

- Check that all users are appropriately qualified and have relevant permissions
- Check that all projects are appropriate and have relevant permissions
- Check that all data access is limited to the minimum participant count and granularity necessary to achieve the analytic objectives to a high standard
- Ensure that all access arrangements are appropriately time-limited
- Check that lapsed or otherwise incomplete projects have their permissions reviewed and revoked

It is useful to map these onto the popular and informative "Five Safes" principles that have been developed as a framework for thinking about safe approaches to working with potentially disclosive data.

Safe projects: is this use of the data appropriate?
The precise ethical and governance rules by which projects are selected as being acceptable or desirable is outside the scope of this review; it is crucial that this is done openly and efficiently, to high standards, following clear and robustly tested principles, with any exceptions to those rules clearly flagged and justified.

Safe people: can the users be trusted to use the data in an appropriate manner?
Checking that users are appropriate is similarly a crucial part of the organisational and governance layer around a TRE. Ensuring that only users and projects with appropriate permissions are able to access and execute code on the platform is similarly vital.

Safe settings: does the access facility limit unauthorised use?
This is covered extensively above.

Safe data: is there a disclosure risk in the data itself?
This is covered above in features ensuring obstruction and detection of re-identification, and where necessary in data preparation using techniques such as minimisation and sub-sampling. Previous work on the Five Safes and TREs has emphasised the important of data being pre-prepared before the analyst is able to access it: this should be done in a way that does not conflict with efficient data management; for example code, written by the analyst to execute against the canonical raw data, can be reviewed for appropriate minimisation or sub-sampling prior to approval for execution.

Safe outputs: are the statistical results non-disclosive?
This is covered above in output workflow.

TREs are not theoretical constructs: they are real, running services. The letter to GPs from Department of Health and Social Care (DHSC) in August 2021 - announcing that NHS GP data would only be accessible in a TRE - stated that this TRE work would build on best practice in existing services such as the ONS TRE and OpenSAFELY. A summary of these two TREs is given below: historically ONS SRS has focused on the service wrapper, and OpenSAFELY on the software layer.

ONS Secure Research Service

ONS is a non-ministerial government department responsible with a range of responsibilities and statutory duties including: collecting and managing a range of person-level datasets including the census, death certificates, and various intermittent population surveys; and producing a range of official statistics related to the economy, population and society at national, regional and local level.

ONS has run the SRS for approximately twenty years as the sole mechanism by which external users can access potentially disclosive pseudonymised person-level data from their own datasets. Data controllers in government (such as the Department for Education) are able to deposit data in the SRS for onward use by others, and collaborative partners such as Administrative Data Research UK have done excellent work encouraging, facilitating, and resourcing data controllers in government to do so.

The ONS SRS and team helped to shape many of the modern norms around a TRE, especially around the service wrapper. The typical user journey is as follows: external applicants request access; this is reviewed and approved; individuals are trained and examined for relevant skills and knowledge around secure data management; the dataset requested is then prepared for the analysts by internal ONS staff, who also use technical barriers to re-identification such as minimisation and sub-sampling of the population; the data is then provisioned, typically into an interactive desktop environment, where users can interact with it using tools such as SQL, Stata and R; a log is kept of keystrokes and mouse movements, to be reviewed where concerns are raised.

The ONS SRS has delivered a very large number of research outputs over the course of its life. During the period of COVID-19 it was additionally populated with a range of health datasets including HES and some derivatives of raw GP data extracted from practices for the purposes of managing the pandemic. The SRS is now expanding into the Integrated Data Platform which will expand the technical features of the service, and meet the needs of internal users at ONS.

OpenSAFELY: an open and secure TRE for RAP

OpenSAFELY is an open source publicly funded TRE created during COVID-19 by the University of Oxford in collaboration with NHS England, LSHTM, and EHR software companies TPP and EMIS. It is currently executing code across 58 million patients' full GP records linked onto other datasets such as HES/SUS, vaccination and death certificates, with a large number of completed and published research outputs in high impact journals.

OpenSAFELY has implemented various technical features to support privacy, transparency, and open science. Analysts use standardised tools for data curation, meaning all code can be reviewed, understood, improved and re-used by other users. These data curation tools also generate "dummy data", bespoke for each project, in the same format as the real research data; analysts use dummy data to write analysis scripts, which are then executed against real data, without analysts ever needing to interact directly with real patient records.

All platform activity is logged and published in real-time. All code for the platform itself is shared for security and scientific review. All technical and user documentation for the platform is openly available online. All code executed against patient records is logged and automatically published when results are published; most code is amenable to review and re-use as it follows standard structures. As a result of these various design choices, OpenSAFELY has strong support from professional bodies, privacy campaigners, and citizens' juries.

OpenSAFELY is portable software, rather than a single TRE, and can be implemented where data already resides: it is currently deployed inside the data environments of EMIS and TPP, where NHS patients' records are already stored, with federated analysis between data centres; and is being implemented in other NHS settings.

Meeting needs for expert and non-expert users

From the long list of user needs listed above, in discussion with analysts and TRE teams, one issue warrants closer review: TREs must meet the diverse needs of different analysts with different levels of technical ability. In essence there is need a for two varieties of TRE, or two windows onto the same underlying TRE infrastructure: a simple model, like a remote desktop; alongside a more complex and flexible model.

Remote desktops are likely the most prevalent form of TRE at present (albeit that the technical features of TREs tend to be poorly documented): the user logs in remotely to an interactive environment that feels much like their own laptop, with familiar tools such as Excel, Tableau, R or Stata, and a specific requested dataset accessible in a folder. This is a convenient way of

working for many analysts, and requires almost no new skills, as it is familiar from the old way of working with downloaded datasets on the user's own machine. It is important that this approach continues to be available, to meet the needs of users who do not have the aptitude or requirement to develop new skills; but it is also important to recognise and mitigate the problems it can cause around privacy, transparency, and open working.

For example, it is hard for a remote desktop environment to detect misuse: more specifically, it is hard to keep and share informative activity logs, as the user is in an environment where they can use tools interactively through a graphic user interface. At most, logs can be kept of keystrokes and mouse movements (a little like trying to store information about a ballet through the footprints left on the stage) or a video of the screen: these are only amenable to slow manual review. It is also harder to prevent data being misused: for example, all the data is readily viewable in an interactive data analysis tool. Because of these shortcomings, additional safeguards are often put in place for remote desktop services. For example, at ONS SRS - a strong and positive example of a good remote desktop TRE, with a very strong governance wrapper - datasets are pre-prepared by ONS staff for each researcher in ways that try to minimise the disclosiveness of the data, by pre-preparing some of the variables, or providing only a random subset of the whole population.

Remote desktop TREs also do not lend themselves to open working, or a network of users collaborating through shared modules of code: users are often not working in ways that meet RAP standards; and exporting code at all can be challenging, since that code has been written interactively while viewing the real data, and is therefore has a higher risk of accidentally containing some disclosive information.

At the other end of the spectrum, there is what might be called a Full TRE, a flexible computational environment that runs in a more script-based manner. This is not an unusual model: indeed, many TREs support it. In some settings it would be the absolute norm: for example, when users run code across high performance computing clusters, there is a natural expectation that most will have the strong computational skills to develop or optimise code and execute it in a largely script-based environment (they will likely also collaborate with, and support, those with strong skills in an adjacent field, such as a specific area of medical statistics).

This way of working requires that users have somewhat deeper skills, such as those in the chapter on Open Working; but the trade-off is substantial. RAP working with shared code is actively facilitated, and indeed is the natural way to work. It is possible to keep informative logs of all activity, and automate a range of efficient working practices around re-used code for core common activities of data curation and analysis. It is also much more straightforward to keep informative logs, share activity logs, and implement a range of security provisions that protect patient privacy by preventing or detecting misuse. Where there are additional privacy provisions, a full TRE can justify deeper access to more detailed and complete data while still maintaining the trust of professionals, the public, and patients. Giving users who wish to work in this kind of environment a remote desktop would not be "easier": in fact, it would substantially obstruct their work.

These models are presented as two extremes to illustrate the different user needs and the opportunities in each broad type of approach; there are also various TRE services that represent a halfway house between the two models, and it is useful for both to exist in parallel: for example, tools and code run in a full, script-based TRE can be used to generate stripped-down datasets, or an easy-to-use interactive service, for provisioning into a remote desktop TRE. At this point it is useful to consider the components of a TRE.

The practical components of a TRE

A TRE can usefully be conceived of (especially when considering construction) as having three components: a service wrapper; some underlying generic computational and database services; and bespoke, subject-specific code for NHS data analysis. These are each considered separately below.

The service wrapper

This is the set of rules, regulations, governance and customer service that surrounds a TRE. There will be a range of rules around who can access the data, the skills or certificates they may need; there will be a similar range of rules around permissioning for projects; there will be processes to evaluate compliance with these rules; there will be forms to collect the data, and administrative processes to manage them; and so on. There will be governance for the TRE as a project in itself, and a range of permissions, contracts, relationships and governance arrangements around the patient data that is being ingested. There will be public-facing material to be managed, describing activity in the TRE to a greater or lesser extent. There will also usually be an "output checking service": when an analysis is complete, and the analysts are ready to release their tables and graphs from the secure environment, this is a final manual check to ensure that no disclosive information is being accidentally sent out. This output checking work is done by staff with skills in data management and analysis: it is a rapidly growing field of work that is long overdue for methodological innovation embodied in re-usable code.

At present, it is clear that there is a wide array of different approaches to this service wrapper. Some are broadly similar; some are very different indeed. Some services that call themselves TREs appear to be much weaker, especially when the service is actually just for internal use inside one organisation. This variation makes little sense, as all TRE services are doing largely the same work. The duplication of IG and governance (and risk of sub-optimal practice) is particularly concerning for those TREs that have been created to house little or nothing more than another duplicative copy of NHS GP and HES data (in itself already creating duplication of infrastructure, risk, and further variation in implementation in a manner that obstructs portability of code). Overall, it is clear that there should be a single set of permissions processes, paperwork, and requirements for each TRE, and a single "recipe" for the service wrapper; with deviations as exceptions where this is genuinely warranted by substantial differences in user need or service need that genuinely cannot be addressed by extension of the standard recipe. ONS has done excellent work in this field over two decades, and developed a very robust approach; SAIL/UK-SERP have similarly developed a standard approach that they offer to external users of their services. This prior work should be built upon.

Generic compute and database

A TRE is a truly multidisciplinary project: it requires strong knowledge of governance, but also deep knowledge of technical infrastructure. However, much of this technical work presents broadly similar requirements to that in other non-health sectors around tasks such as provisioning large databases, ensuring that they are secure, ensuring they are performant, managing access permissions, and providing a computational environment where users through some sensible means can call up processor power, memory and disk storage to execute their code.

A TRE is a truly multidisciplinary project: it requires strong knowledge of governance, but also deep knowledge of technical infrastructure.

There is some overlap, inevitably, with domain knowledge: the team ingesting and transforming the data will need some knowledge of the specific qualities of the data they are working with, and the end-users' needs; the implementation of processes around users will entail some knowledge of the service wrapper processes. But overall, the key practical issue is as follows: the compute, database, and service design aspects of a TRE are largely generic tasks that can be largely delivered by staff who have strong generalist software and data science skills, and who are easily recruited from other sectors. This is vitally important, as there is a drastic shortage of people with strong knowledge of both software development and the specific challenges of NHS data. Furthermore, individuals with strong skills around the delivery of data infrastructure in other areas are likely, in many cases, to be better placed to guide the development of this aspect of the work - or monitor work by others - than most senior academics in epidemiology, or NHS analysts.

Subject-specific code

The work above will deliver a raw, powerful, general purpose computational data science environment. From discussion with technical teams who have provided such environments for health data in England, and then found that experienced epidemiologists were unable to use them, it is clear that this work is not sufficient in isolation. There is then a need to bridge the gap between a general purpose computational environment, and everyday health data analysts with statistical coding skills in tools such as Stata or R, but less experience with data science, software carpentry, and efficient pipelining for extremely large volumes of data. This is not an unusual problem to encounter, and belies an important truth for all ambitions to create robust TREs: a full TRE for electronic health records analysis is not a product that can be procured, but rather a service that needs to be built, in interaction with real users.

This is best understood by analogy with other fields. The data science team at the music-streaming service Spotify do innovative work with data that helps drive the usability and popularity of their subscription service. For example, they extract patterns in the listening behaviour across all their users, and then use this to provide individual users with tailored recommendations for other music they might enjoy. The Spotify data science team couldn't buy, off the shelf, a data science environment specifically built to service the needs of "a global music streaming service". They implemented standard off-the-shelf tools for a general purpose data science environment. Then, within that raw environment, they needed to build their own tools, analytic approaches, workflows, data preparation work, and so on. A new arrival in the Spotify data science team today will find modules of code, libraries and packages - some even with nice interactive interfaces - to help them find interesting new patterns in Spotify user data. Many of these tools will feel like part of the furniture, but they were all built by their predecessors in the Spotify data science environment. Furthermore, many of these tools will not have been built to a pre-determined specification, by software developers hired to do that work to order; rather, they will emerge from a team. A single analyst might painstakingly implement a one-off analysis; if it looks like the approach will have broader use, then a more experienced developer might offer to help package it up into a function or library, with good documentation; if it becomes a commonly used approach, they might work with other analysts to create an interactive tool.

This is a normal challenge for any community of data users to address, in any sector: analysts and developers collaborate to meet the challenge of developing bespoke code and working methods for their own bespoke needs around data management and data analysis, to solve their own business analytics problems. They work within general purpose data science environments, turning their work into shared code and tools for others to use as they go. Building national and local TREs for NHS service analytics will only be possible by following the same approach. It will require a team of developers working alongside real NHS service analysts and researchers to produce bespoke functions and code libraries, new working practices, and up-skilling users.

The line between TRE code, and user-generated code

In this sense, there is also a somewhat blurred line between TRE code produced as a core resource, and the code written by users to run inside a TRE, especially when code produced by one user meets the needs of many others. If one analyst produces an elegant way to create an "ex-smoker" variable for each patient, running against the raw patient data accessible in a given TRE, then others can re-use it as a single command. If one researcher produces an elegant python function, created to operate against the resources inside a given TRE, to do "case matching" (a common task in epidemiology studies) then others can re-use it. If a team produces a robust pipeline to produce interactive graphs of service activity, created to operate against core data and code resources inside a given TRE, then all other users can use and re-use it.

There are almost limitless examples of regularly performed tasks that lend themselves to a standalone function or library of code in a TRE for a wide range of users. A core team might produce a "point and click" graphic user interface that can help less-skilled users specify a dataset request in code that can execute in the TRE. A similar team might produce an easy window to generate graphs of service activity volume over time, broken down by organisation and month; or to do the data preparation work, so that analysts can do the final aspects of the work themselves flexibly in Tableau. A team of security researchers might use a TRE to run research analyses on "uniqueness" of individuals in datasets, and develop or evaluate methods to obfuscate disclosive personal data; and then turn this theoretical work into re-usable code that can be used to automatically check the extent to which data has been adequately minimised before provisioning into a remote desktop TRE.

This is the kind of mixed collaborative and competitive ecosystem that can arise spontaneously in other settings, but requires secure infrastructure and modest coordination in the NHS, only because the underlying raw data is so sensitive and disclosive. When TREs are used to resolve the conflict between broader access and privacy then this new way of working is unlocked: it creates the environment that can host a modern, open, collaborative ecosystem of users, spread across multiple organisations across the public and private sectors. It lends itself immediately to open code from publicly funded work; or even (with more effort around the relevant infrastructure) an "app store" model for privately developed code. There is no need for all

this additional code to be built overnight: the key is to build the smallest possible number of TREs, and produce reasonably consistent computational environments where all NHS data users can work in the same way, with the possibility of sharing and re-using code. If this is done, then the NHS and health research sector can replicate the productivity and efficiency around data seen in other sectors, and avoid the duplication and closed working that have become the historic norm for NHS data management and analysis.

Diverse datasets, diverse users

Before considering the best model for TREs it is useful to give brief overview of commonly used datasets, and common users of data, as each can present slightly different challenges and opportunities. From extensive discussion the majority of datasets and users are captured in the following categories.

Users

The following list is provided as a guide following extensive discussion with the community: many individuals will sit in roles that cover two types of organisation.

National NHS service analysts

These reside in multiple different teams, in multiple different settings, and large organisations such as NHS England have multiple analyst teams, sometimes with little formal interaction, each maintaining their own local copies of key datasets, or derivatives of them. They typically work on national datasets, sometimes on registry data (often in collaboration with external groups), and often on producing or using organisation-level datasets (such as “Model Health System”).

Local NHS service analysts

These can reside in CSUs, CCGs, GP Federations, AHSNs, Trusts, local authority public health intelligence teams, and more recently ICSs. They work on national and local NHS datasets.

Commercial providers of analytic services

These are typically commissioned by local and national NHS organisations. They often maintain their own large databases of extracted NHS data.

Academic researchers

These typically work in universities, and typically access NHS data by downloading it onto local machines.

Innovators from the life sciences sector.

Companies making treatments sometimes access data for research into the benefits and hazards of interventions by downloading it for their own teams to analyse, or work in collaboration with academics. “Digital health” innovators often struggle to access NHS data at scale.

Regulators

They access various datasets alone or in collaboration with others.

Government analysts

They are typically based in DHSC or similar departments. In the past it was common for these teams to be closely engaged in service analytics using NHS data; over the past two decades their access to data has greatly diminished, although there is a strong appetite for better access to do more work.

Datasets

Notably there is no clear single map of NHS datasets (albeit that there is also substantial doubt from many in the applied data science community, including the Director of the Open Data Institute, about the value of simple catalogues for such data). The following list includes only datasets containing person-level data, and mostly “event-level” data; it does not include the many datasets that are solely about organisations such as GP practices or hospital trusts.

Commonly used national NHS datasets

These include GP data, Hospital Episode Statistics (HES)/Secondary Uses Service (SUS), prescribing data, and other similar datasets with national coverage. They are typically extracted from local or national organisations, and held nationally by NHS Digital. These datasets are very widely known and used. These datasets will provide the greatest insights, at the highest speed, and will deliver the biggest return on investment for TRE work to make them more accessible.

Bespoke local health and social care datasets

These types of information are typically collected by local NHS analyst teams, often working in (or with) Commissioning Support Units (CSUs), but also in Clinical Commissioning Groups (CCGs), Academic Health Science Networks (AHSNs), Trusts, local authority public health intelligence teams, and now Integrated Care Systems (ICSs). Data includes extractions of very detailed patient data from local hospital systems in a diverse range of formats, to varying degrees of granularity for different hospitals and local organisations; social care data in a very diverse range of formats (such as invoicing data for care activity from a local authority). However much of the work by local analysts in these setting will also be conducted on smaller local cuts of the national NHS datasets such as GP data, or HES/SUS. These local datasets are an under-used resource: if they can be made securely accessible in a TRE for both researchers, commercial vendors, and service analysts on an equal footing with a requirement for open working methods by all, this will create a rich, open, collaborative ecosystem of data use for service improvement and high impact research.

Registries and audits

These are typically based around an extract of routine EHR or NHS data, such as HES, combined with some additional bespoke data collection: for example, some information from a bespoke online data collection form filled out by clinicians or local NHS staff about the brand, serial number and location of an implanted device; a detailed description of a hospital procedure; or a detailed account of inpatient treatment for a given medical condition. They typically cover a single topic or disease area, such as the National Cancer Registry, or the Renal Registry. These are sometimes run by NHS or public health organisations, sometimes academic groups, and are often mixed between the two. Some are collaborative resources that are open to use by a wide number of individuals and organisations, some are closed projects where only a small team can access the data, and many fall somewhere between these two arrangements. Some are very large, some are very small, and many are administered in small single topic organisations, in extremis on one individual’s computer. Some of these projects have modest personnel and resource: they are often a “labour of love”. Many have fallen “beneath the radar”. Overall, this area is ripe for rapid innovation and more efficient data use in RAP environments with shared code and broader access. A key challenge is that current resourcing for these projects often regards all activity as a package, where one organisation is resourced to deliver the entire chain of data collection, data management, and production of final reports and academic papers, being judged largely on the latter alone. Making

all these datasets into a service available for wider use in TREs will therefore require that the core work of data collection and management is adequately resourced and rewarded.

Single research datasets collected by academics

These can vary widely in scale. At one extreme there are small single bespoke data collections for individual studies, which can sometimes have some re-use value. At the other extreme are the "cohort studies": these are large bespoke research data collections where the same group of people are variously interviewed, tested, scanned, bled, and otherwise examined at regular intervals over the course of many years. Some of these projects have been running for many years, including the 1946 birth cohort, which has regularly collected data on around 3,000 currently living participants since their birth over 75 years ago. These data collections support a wide range of data analysis projects. As with the registries and audits: these cohort studies are often a "labour of love" for a small group of individuals; some are closed projects where only a small team can access the data; some over the past decade have won funding that allows them to support multiple external users and act as a common research data asset. These cohorts have often struggled to get permission to access NHS data to match onto their research records. These datasets represent a key national asset; there is strong support in the cohort study community for their data to be used as a national resource where there is funding to make this practical; delivery of secure shared analytic environments will secure this outcome.

Other datasets

Alongside these there are various other less well known or documented datasets, some national, some partial or whole derivatives of the above, but many supporting a key function somewhere in the NHS or research landscape: they are noteworthy because they emphasise that there will often be edge cases for national policies around data management.

Overall, while rationalisation and de-duplication is important - to manage risk and contain costs - it is clearly not possible to meet all the needs of all users, and all these datasets, in one single national TRE.

Current Trusted Research Environment and Data Access Environment provision

Much research work with NHS data is done by individuals downloading pseudonymised patient records onto their own computer. Alongside this there is also a long history of investment in various environments for data analysis in the NHS and academia, with some mixed NHS/academic projects. From discussions and desk research it is clear that this has been patchwork and typically resulted in closed "black box" projects with limited sharing of technical information; nonetheless in these projects and others excellent work has been done by well-motivated and talented teams in diverse settings.

National provision

Nationally, NHS Digital has had a longstanding plan to build a Data Access Environment and move the majority of analysis on national datasets to this setting. More recently, work has begun on a full TRE with appropriate service wrapper and scale. The Health Data Research UK-British Heart Foundation "Data Science Centre" is an example of an early tenancy in this NHS Digital TRE. Since the announcement that the planned GP Data for Planning and Research data collection will be TRE only, there are plans to substantially enhance NHS Digital's TRE work in order to serve the needs of all analysts working on national GP data at scale. Other national organisations working with data at national scale tend to have their own internal data analysis environments, with varying implementations, and sometimes different arrangements for different teams working on the same or similar data within the same organisation.

Local NHS provision

For local NHS service provision, there is a complex patchwork of arrangements covering different sizes of area including administrative organisations such as CCGs, PCNs, Federations, Trusts, and now ICSs; Commissioning Support Units supporting a wide range of organisations; and diverse wider regional arrangements such as Local Health and Care Record Exemplars and standalone projects such as Connected Cities North. These projects are very variable in scale and sometimes entail data analysis work as a project sitting alongside a separate project to create some form of linked dataset of health and social care records that has been principally produced to inform direct patient care (for example, making some aspects of patients' records from one service available to clinicians or staff using computers in other local services). These projects are very diverse in nature and implementation; many are closer to being an internal DAE than a TRE. Various service analysts and academic researchers expressed concerns to the Review team that projects containing data covering their region, that were relevant to their work, were not accessible to them. With notable strong exceptions of open working, typically there is very little accessible public information about the technical implementation or service wrapper of these DAE or TRE projects, other than PowerPoint slides or public relations material. While this is likely to reflect contemporaneous norms and the expectations of commissioners, it does substantially limit the scope for learning from their experiences about the best approaches to rationalise, standardise, or create tools for federated analytics across these projects.

Academic provision

For academic TREs there is a very challenging patchwork of DAEs and some projects that could be considered a TRE, alone or in collaboration with NHS partners. Some contain NHS records, some contain data collected for research, and some contain both. Some of these are constituted as projects where data is accessible to external users; however, in some cases users raised concerns that they had found that access was, in their view, unreasonably withheld to these shared resources. Overall, there is a tendency for DAE and TRE projects to be regarded as "black box services" where external visibility is largely limited to public relations material or PowerPoint slides, and completed single research papers, rather than code or technical documentation for the TRE itself. Again, in many cases this is likely to reflect prevailing norms and funder expectations rather than individual choices. Various members of the research and policy community expressed the concern that they sometimes found it hard to see or understand what had been spent, and what had been built, with various academic TRE investments. While there have clearly been some excellent and productive environments produced, there was also a strong reported sense by some of substantial plans being set out, apparently resourced, but then not delivered, only to be replaced by new plans. The review team attempted to conduct a rapid informative overview of recent investments made - in particular during COVID-19 - by research funders on TREs and related projects, including code and methods for data management, secure analysis, and similar tasks. It is not the role or objective of this review to produce a critical public audit of individual projects; however, overall, with some very positive exceptions, this overview proved challenging, even with direct questions.

SERP and SAIL: shared TRE resources in Wales

The Secure Research Platform (SeRP) is based at Swansea University and led by Professor David Ford and Professor Simon Thompson. It grew out of the need to store data for the SAIL Databank in Wales, which was created in 2005 to curate and provide access to all public sector data of Wales. Recognising the need for secure storage and analysis of data across the wider UK community, in 2011 the team created SeRP as a general purpose environment, to make it possible for multiple, complex datasets to be linked, managed, analysed and shared in accordance with data provider permissions and appropriate Information Governance requirements.

SeRP is a customisable analytics platform with a range of features that can be tailored to suit a particular data owner's needs. By using SeRP, tenants - other external teams and organisations - can store and link their detailed research data, health records or other datasets inside SERP's cloud infrastructure, manage it and then provide secure virtual access for analysis to their own team, but also to others in the wider research.

As an example, SeRP now hosts ALSPAC, the Avon Longitudinal Study of Parents and Children (also known as Children of the 90s), a cohort study collecting detailed information on 14,000 pregnant women and their children over the course of three decades. The data is stored in SeRP, but used by ALSPAC analysts and others to deliver insights into a range of social and clinical research questions.

Researchers are able to utilise a wide range of analytic tools on top of underlying raw data, by making use of SeRP's "off-the-shelf" components. In addition to the customised environments deployed inside SeRP, researchers can also access shared project spaces within the research environment to enable wider collaboration through database space, file store, meta data libraries, concept library, wiki, Git and other support and help materials. SeRP offers a wide range of computation environments, including GPU and HPC clusters as standard.

SeRP offers a broad approach to governance, building on its ISO 27001 and Digital Economy Act accredited status, helping to reduce duplication of effort and needless variation in implementation. SeRP operates as a private research cloud with an interface that allows non-technical members of the research team to control access to various resources within each project. With over 35 tenants in the UK and internationally, over 3000 users and many hundreds of ongoing projects, SeRP is a strong example of researchers working collaboratively to develop shared resources, and a strong example of a TRE that has met the needs of a wide range of users over a long period of time, producing a diverse range of completed analytic outputs. It has also recently begun to share more code and technical documentation, helping to lead on the development of this emergent norm in health data research.

SeRP is a strong example of researchers working collaboratively to develop shared resources, and a strong example of a TRE that has met the needs of a wide range of users over a long period of time, producing a diverse range of completed analytic outputs. It has also recently begun to share more code and technical documentation, helping to lead on the development of this emergent norm in health data research.

Recommendations

All analysis of NHS patient records should move to be done in a TRE. This will allow more users to access NHS data while preserving patient privacy. It will reduce duplication of risk, work, and cost. It will also help to drive the overdue move to modern, open working methods and RAP.

The national TRE strategy should aim to:

- Address privacy needs proportionately
- Minimise duplication where possible
- Maximise open working and RAP for productivity and quality
- Avoid re-creating monopolies and siloes
- Be realistic, with step-wise progress

In general work should be prioritised on the following basis:

- The most commonly used datasets
- The most detailed and disclosive datasets
- The datasets with the largest population coverage
- The datasets with the most potential to improve patient care, if more people could safely access them.

National TREs

Large disclosive national datasets should only be available in a national TRE, even when they are pseudonymised. This should begin with the GP data but expand to cover other disclosive datasets over time. Any special case exemptions should be resisted as they are likely to repeat the strategic shortcomings of the past; however, the justifications for these proposed exemptions should be well-received and used to improve the TRE model. The number of national TREs should be kept to the smallest possible number, but with the possibility of expanding to three in total, only to mitigate key risks: non-delivery; and monopolies around access. Every TRE containing national NHS data should be a shared resource where all users can apply for access including national NHS service analysts, local NHS service analysts, academic researchers, and others. All TREs should support and require RAP working methods.

Local TREs

ICSs are new organisations in the NHS, all using data to improve the quality, safety and efficiency of care. They are the perfect opportunity to rationalise local TRE provision. All local TREs for ICSs should ultimately conform to a single national model, with pragmatic flexibility to account for diverse local datasets. All analytics in these settings should be delivered with RAP methods. All Local NHS TREs should have a common TRE Service Wrapper. Local NHS TREs should be able to request their patients' NHS data, cut from national datasets, to be transferred to them as needed.

Academic TREs

Academic TREs and DAEs should be recognised as a challenging and, historically, often closed space. All academic work on NHS patient records alone should always be conducted in NHS TREs, and compliant with RAP and open working methods. Patient data should only be transferred out to other non-NHS data centres when that patient has consented for this to be done (for example in consented clinical trials or research studies). There should be strong direction to create standard technical and open software for TREs containing academic data to ensure code is portable, shared, and work can be federated, as per NHS TREs. All funding for academic work on TREs should pass through a single national organisation, with clear accountability into government. This funding coordination body should publicly disclose, for every project: the amount; the timescale; the funding source; the target organisation and team; the core objectives; and the location where updates and open delivery can be seen. The majority of funding

should be competitive and open to all. Academic funding should be principally focused on delivery of innovative methods and working open code with adequate documentation. All academic TREs should use a common TRE service wrapper modelled on the standard NHS TRE service wrapper.

All TRE work in all sectors should be approached as an open service, driven by open code with adequate technical documentation. Leadership should come from those with appropriate technical skills and proven delivery on data infrastructure platforms. Below are a range of detailed recommendations to ensure efficient delivery and address previous challenges.

Delivering a national TRE programme

National TREs are a substantial piece of work with an extremely high return on investment. This work is currently in the planning stage within NHS Digital. This must be a high status and well-resourced project delivering a core piece of NHS infrastructure, not a small side project. Delivering a national TRE for the GP Data for Planning and Research data collection will do all the groundwork necessary to create a strong, performant, general purpose TRE, as the GP data poses a larger computational and data management challenge than any of the other smaller national datasets.

TRE1. Create the role of National Lead on NHS TREs and Data Curation

This individual should have strong generalist CTO skills and experience of coordinating data infrastructure projects in any sector; they should be trained in strong domain knowledge of NHS data and analytics after their appointment. They should have no responsibility whatsoever for any single specific analytic output using NHS data: their focus should be solely on TREs and data curation, leading and delivering the national NHS TRE Technical Delivery Team and holding responsibility for the work set out in this section. This role can sit in the NHS Transformation Directorate but must directly supervise the teams delivering TRE and curation work. They should be supported by a team that includes a national lead for creating a robust governance and service wrapper around all TRE activity.

TRE 2. Rapidly create a substantial multidisciplinary TRE technical delivery team

This should be rapidly convened by the NHS Transformation Directorate. It is crucial that this team has the right people: it should combine skills in software development, data architecture, clinical informatics, data science, data management, information governance, cybersecurity, and open source software. This work cannot be done by intermittently consulting people with domain knowledge and technical skills as external advisors; these skills must be strongly represented on the core project team as internal staff. Hands-on NHS service analysts and academic researchers with strong skills in some of the prior listed domains must be closely involved in the core team from the outset, but only as expert users: the project must be led by software developers and technologists, not researchers. This team should seek out and involve, by secondment or close advice, the best teams nationally and globally with a proven record of successful completed delivery of TREs in health and other domains including: ONS Secure Research Services and OpenSAFELY as per Secretary of State's letter to GPs; along with UK-SERP, Genomics England, Public Health Scotland, and others. It is important to cautiously avoid, from this expert group, teams who have had resource to build TREs but not yet displayed any public code, outputs, or technical documentation; or organisations who have simply been intermediaries for funding to other organisations who have themselves built the TRE tooling.

TRE 3. Rapidly agree and publish features for the minimum viable National TRE

This team should agree and publish within 2 months, then finalise within 2 months, the core technical design features of a strong MVP for the National TRE that can meet all use cases for the national GP data extract. This MVP must support RAP working, and have a robust service wrapper. It should ideally have underlying compute infrastructure that can scale for a full National TRE.

TRE 4. Agree and publish proposed features of the full National TRE

This should be a flexible framework that is open to iteration as new user needs are identified, and as experience is gathered from delivery of the MVP.

TRE 5. Produce a minimum viable TRE in 6 months

Examples of national TREs for NHS data already exist, so MVP implementation that is capable of scaling, and demonstrating rapid delivery, should take no more than six months. This should include a full secure data management and analysis pipeline, most core features, code executing against live data, sharing code as per RAP, and delivering a small range of completed outputs from a range of user groups. Initially this should be delivered as: a performant service wrapper; underlying core services such as database and provisioning of compute; and a range of 2 to 4 TRE service options (a remote desktop, and code execution environments) operating as software layers that use these underlying core services. This de-risks the project as it reduces lock-in to any one TRE platform option, incentivises delivery, permits iterative improvements, and allows the best features of each system to be used.

TRE 6. Rapidly scale over 18 months

The following 18 months should be spent onboarding external users, iterating the platform, iterating detailed documentation for users, actively recruiting users, and adding features in response to user feedback. Proof of delivery should be in the form of outputs, code, and technical documentation, not PowerPoint slides or communications material. The team should aim to support the delivery of at least 100 completed and publicly available end-to-end analyses of real NHS data, that meet genuine NHS service or academic research needs, from ten different teams of users, with the data management and analysis code shared openly with adequate technical documentation, by the end of year one. There should be at least 10 python functions, or similar, available on GitHub, that meet core common user tasks within the TRE, with more than one user. There should be full technical documentation of all platform features, and underlying datasets, openly accessible and associated with the relevant underlying code.

TRE 7. Include GP data and certain commonly used national datasets from the start

The first iteration of a national TRE should contain the key, commonly used national datasets including GP data, HES/SUS, and prescribing data. GP data is by far the most challenging dataset to support; the others represent a marginal increase in work. This data alone will represent unprecedented depth and breadth of NHS data, supporting a broad array of innovative outputs. NHS service analysts have not previously been able to work systematically with GP data. This alone represents a phenomenal opportunity. Early outputs are likely to include: work evaluating variation in service activity and clinical outcomes between different organisations and regions; monitoring the resumption of clinical activity

following the COVID-19 pandemic; identifying opportunities to improve the quality, safety and efficiency of prescribing; understanding and predicting demand based on detailed data about local patients; and so on. When linked to HES and ONS death certificate data at national scale this resource becomes even more productive, as analysts can create a clear picture of the full patient journey through services. Combined with RAP working methods, to ensure work is systematic and reproducible, this will be a new dawn for NHS data science.

TRE 8. Expand the National TRE in time to accept bespoke datasets

The Minimum Viable Product for the National TRE must be to support the core national datasets. However, there are many external datasets - such as the cohorts, registries and audits - that rely on ingesting large volumes of national EHR and NHS data, sometimes without patient consent. These projects should ultimately move into a national TRE. Doing so will require that a national TRE is capable of ingesting data and supporting analysis across various diverse datasets and data structures; this is readily achievable and should be piloted with a small number of pioneer registries, audits, or cohorts after the MVP is delivered. This flexibility is readily deliverable, and has been delivered in other settings, including ONS SRS, OpenSAFELY, and SERP. However, there should not be an expectation that a national general purpose TRE would be immediately capable of importing very large and complex multimodal datasets (such as genomic data) albeit that they may import and link more sparse derivatives of such data. Similarly, there are likely to be a range of social care datasets in local NHS TRE settings; these will be bespoke to each provider and region; however they are likely to become more nationally harmonised if the recommendations in this review are followed; these similarly are an important dataset to consider ingesting into a national TRE.

TRE 9. Evaluate new developments in privacy engineering; adapt accordingly

The use of TREs is likely to remain best practice for the protection of disclosive NHS data for some time to come. It is however reasonable to expect that the core technical design features of a good TRE will shift as technology develops. The team should annually review the core technical features expected of national and local NHS TREs to ensure that NHS TRE provision does not fall behind best practice for privacy preservation.

TRE 10. All TREs must support code sharing and RAP

Adoption of modern open working practices is a core benefit of moving to TREs. All TREs at all scales must support code sharing and minimum RAP working methods.

Develop Trustworthy, Agile, Standard Governance for National NHS TREs

TRE 11. Build a TRE governance team to create a robust framework around TRE access

A good TRE must be surrounded by good governance, and support this with relevant technical features. This team must contain experts in information governance, policy, customer workflow, and TRE implementation, with close involvement or direct inward secondment from other TRE teams at for example, NHS Digital, Office for National Statistics, Genomics England, and UK-SERP. This team must be tasked with building processes to ensure that: all users are appropriately qualified and have relevant permissions; all projects are appropriate and have relevant permissions; all data access is limited to the minimum participant count and granularity necessary to achieve the analytic objectives to a high standard; all access arrangements are appropriately time-limited; all lapsed or otherwise incomplete projects have their permissions reviewed and revoked; all projects appropriately involve patients in their design; and any additional aspects of work they identify in the first 2 months. This should build on the best prior workflows in for example, NHSD, ONS, and SAIL/UK-SERP. The team must be specifically tasked with ensuring that access is swift: specifically, they should be required to report back formally to senior leaders every month for the first 6 months on any barriers to fast platform access - whether these are regulatory, legislative, technical, practical, resourcing or organisational - so that these barriers can be rapidly addressed.

TRE 12. Create a single standard Service Wrapper model for NHS TREs

This should cover issues such as safe people and safe projects; a standard approach to ethics and governance; and a standard approach to Information Governance (IG) and onward access arrangements for datasets ingested into a national TRE.

TRE 13. Create a national standard approach to "output checking" and support automation

All results tables and graphs leaving TREs under best practice must currently go through manual "output checking" to ensure no disclosive material is accidentally released: this is time consuming and implementation is variable. The National TRE team should develop a single national standard on best practice for output checking, then require and monitor adherence. They should also collaborate closely with academic teams and funders aiming to automate aspects of this work (as per the list below of illustrative funding priorities for data science infrastructure).

TRE 14. Establish a standard scheme to accredit NHS TRE users

There is a need for a single accreditation framework to identify that individuals have appropriate organisational credentials and appropriate training to work safely with patient data. This should mirror the accredited researcher scheme run by the ONS under the Digital Economy Act (2017). Researchers wishing to access data via any NHS TRE should become accredited NHS researchers. Once accredited, researchers should be able to use any of the licensed TREs for approved research projects without having to apply individually for access.

TRE 15. Ensure TRE access is faster and easier than data dissemination

TREs are lower risk than data dissemination. They present a range of safeguards that substantially increase patient privacy, and prevent analyses outside of those permitted, by sharing information about all activity in the platform, and by using tools that obstruct and detect misuse of data. The current IG arrangements - widely regarded as slow and obstructive - were developed to manage the risks of data dissemination. They should not apply to TREs. As discussed in the IG chapter, TREs should be subject to a lighter touch regime that is proportionate to their lower risk. This will incentivise the use of this safer and more open approach.

TRE 16. All TREs must share live detailed activity logs

These should openly disclose, at minimum: the individual executing code; a link to the documentation for their legal basis to process the data; the datasets against which code is being executed; and ideally the data management and analysis code itself, as this provides the clearest and most unambiguous record of

what has been done. This need not include the results of the analysis; only the code, and ideally accompanying documentation, as a public record of what has been done. This will allow the system to maintain public trust that users are only conducting appropriate analyses with their NHS patient records.

TRE 17. Create clear rules for undeclared analyses in TREs

A concern has been raised that sometimes there is a legitimate need for certain users to conduct certain forms of data analysis discreetly, without openly declaring their activity at the time. This is a reasonable user need in certain circumstances. However, this TRE activity should be logged as usual, and published at a later date; appropriate rules around undeclared analyses and delayed disclosure should be created by the TRE governance team.

TRE 18. Switch off data disseminations, without undue panic

There is no need for potentially re-identifiable disclosive data at national scale to flow outside of secure environments. TREs are the only way to build public trust and get larger numbers of people working on NHS data for innovation, service improvement and research. No new data disseminations outside a TRE should be established. All current large-scale disseminations of GP and other granular patient data - both national and local - should be reviewed within six months for replacement with a TRE option. Any needs that are claimed to fall outside of TRE usage should be published and considered by the TRE technical team for iterative development to meet the proposed shortfall. Despite the public and professional concern raised about the GP data extraction, there should be no panic or repeat of what happened in 2013 following the suspension of Care.Data, when several unrelated data flows from NHS Digital to research users (such as HES) were suspended or delayed with no alternative plan for access. TREs are needed to meet the new risks of more detailed GP data, and wider access to data; they also address the longstanding shortcomings of pseudonymisation and dissemination; but there is no new emergency.

TRE 19. Conduct an annual access audit

The TRE Technical Team and Service Wrapper Team should conduct a collaborative annual "audit for improvement" looking at TRE access request and approvals, the count of completed outputs from each TRE, the extent of code sharing and technical documentation, adherence to national standards around service wrappers, and similar outputs.

TRE 20. Publish all technical steps taken to prevent and detect misuse of data

All TREs must take technical steps to prevent and detect misuse of patients' data, such as by obfuscating access to raw data; and disclose these technical methods openly.

Ensuring National TREs are Accessible, and Used

TRE 21. The National TRE should be open to all legitimate users

Any national TRE containing national NHS data should be open to all legitimate applicants following a rules-based system including: national NHS service analysts; local NHS service analysts wanting to work on their own local data in a national context; academic researchers; government analysts from outside the NHS; and, following appropriate and positive consultation with the public, users from the life sciences sector.

TRE 22. No special cases for working outside a TRE

Since the announcement that the planned GP Data for Planning and Research data collection will be TRE-only some organisations have been arguing that they should be regarded as an exception. There is no reason why a TRE cannot meet all users' needs. Any organisation raising concerns - whether due to technical or governance issues - should be fully listened to, their concerns fully understood, and addressed in the design of a national TRE. Any organisation hoping to build a closed data analysis environment for their own internal use should be encouraged to support and facilitate work on a full TRE so that it meets their needs; or to deliver a full TRE themselves as a piece of national data infrastructure. The biggest challenges in delivering complex national data infrastructure are technical: any group that can deliver on that challenge to produce a closed internal Data Access Environment can be supported by those with generalist and governance skills to add the service wrapper needed to produce a TRE. Any relaxation of this approach is highly likely to result in repeated duplication of work and risk, reduced standards on governance and transparency, obstructions to open working and re-use of code, and monopolies around access that do not reflect the needs of patients or the wider system through arbitrary rules - or arbitrary application of rules - around access to data.

TRE 23. Ideally one national TRE, never more than three

Ideally there should be one national TRE. This will minimise duplication of effort around the IG service wrapper, the underlying technical infrastructure, and cybersecurity risks, and minimise unnecessary variation in implementation that has historically obstructed re-use of code. However, in pragmatism, the system should accept the possibility of having very slightly more than one national TRE, only insofar as this is necessary to avoid the risks

of non-delivery and inertia caused by a single monopoly provider, and to stimulate competitive innovation and creative diversity. It is vitally important that the number of national TREs is kept to the smallest number possible, and there should never be more than three TREs containing national scale NHS data. TREs containing data of this scale and depth should only be delivered by national government organisations, ideally within the NHS, with clear lines of accountability into government. Consideration should be given to closing any national TRE that is not delivering where others are.

Two principles should be followed whenever considering any new national TRE. Firstly, a new TRE should only be considered where it genuinely offers new features. Any proposal for a new TRE should be required to demonstrate with substantial evidence that they will deliver new features, meeting genuine unmet user needs; and that they have tried to deliver this additional functionality at similar pace by adding additional features to an existing TRE, but found it to be impossible. This will block needless duplication. Secondly, any TRE must be a shared environment open to all legitimate users. No organisation should be permitted to create a "nearly TRE" for their own internal use. Any TRE containing national NHS data must be a shared national resource where all NHS and other users can apply for access on an equal footing.

Local NHS TRE provision

TRE 24. Create a Local NHS TRE Programme

This programme should have a single clear leader and coordinate work on local NHS TREs ensuring that they follow a common governance wrapper, and a single common open technical model to facilitate portability of code, staff, analyses, curation, and (where needed) federated analytics.

TRE 25. Work to rapidly standardise local TRE and DAE provision, starting with ICSs

Currently local data analysis work is conducted with highly variable working methods, in highly variable computational environments, with even the same national datasets stored and used in very different ways. This is not informative or fertile diversity: it is largely hidden, and happenstance. It obstructs sharing of code, methods, curation, and learning.

This problem can be readily addressed by the following actions:

1. Create a standard service wrapper model for local NHS TREs
2. Ensure all ICSs use a standard TRE approach
3. Encourage other local NHS data centres to use the same standard TRE
4. Manage diverse local datasets by creating and sharing standard data curation tools and methods
5. Ensure all local implementations of national or commonly used datasets such as SUS/HES conform to a single standard
6. Ensure all datasets extracted from national datasets in NHS Digital are requested using standard data management code
7. Ensure local analysts use a national TRE wherever possible
8. Work towards federated analytics with standard local TREs
9. Listen carefully to local NHS analysts and TRE managers who describe shortcomings in standard approaches; and address these wherever possible.

TRE 26. Create a single Service Wrapper model for local NHS TREs

Based on the work for the National TRE, a standard service wrapper should be created by the same team - in consultation with local NHS and academic TREs - to cover access to work on data in these environments. All local and academic TREs should be obliged to use this standard service wrapper for access, to avoid needless duplication of IG activity, and to avoid (or at least document) activity that may lead data access monopolies. Exceptions to the Standard TRE Service Wrapper should be possible, but to a prespecified set of criteria. Exceptions should be publicly disclosed, alongside the justification, under a robust exceptions framework. Any exceptions and modifications made in any local or academic TRE should be reviewed at six monthly cycles by the team to consider whether they justify a revision of the Standard TRE Service Wrapper. The data ingestion elements of this local TRE service wrapper should aim to impose some standards on the current highly variable and expensively duplicative array of approaches to IG for local data flows. All local NHS TREs should recognise the standard accreditation for NHS TRE users.

TRE 27. Ensure all ICSs use a standard TRE approach

Integrated Care Systems (ICSs) are new organisations in the NHS, and are now the primary locus of work for local planning and delivery of care across a region. All ICSs are expected to use data to improve the quality, safety and efficiency of care. This is an outstanding opportunity to drive change. All local TREs for ICSs should be required conform to a single national model of TRE, rapidly developed, with pragmatic flexibility to account for diverse local datasets; then all analytics in these settings can readily move to conform to RAP. The National TRE technical team should rapidly review current planned or actual provision of data analytic services in ICSs, and identify the best opportunities for harmonisation consistent with the principles above. This is likely to entail: a standard open source TRE approach created for ICSs; standard tools (in the formal sense of functions and libraries) for data management to ensure best-case sharing of code, methods and documentation for data curation where a standard TRE is impossible; but not necessarily a standard underlying compute infrastructure. The system should strongly resist the urge to believe that creating a "diversity of approaches" will allow the best model to emerge and spread: previous experience shows that local data aggregation programmes and data analysis environments tend to be closed, black box services with little or no technical documentation from which others can learn, and only arbitrary variations in approach.

TRE 28. Ensure any other local NHS TREs use the same standard TRE approach

There is currently a diverse array of local NHS data aggregation projects as legacy from a range of prior commissioning choices. Many of these are delivering strong service; some may warrant further review. All should be strongly encouraged to adopt a single national approach of a standard NHS local TRE that supports RAP and conforms to a standard NHS TRE service wrapper. This will help to address concerns expressed about transparency, open working, access barriers, and duplication of work. Mixed NHS and academic projects built principally or wholly around NHS data should fall under NHS control. Non-standard TRE or DAE projects containing NHS patient records should be reviewed annually, following commencement of an NHS TRE programme, to establish whether they add value to standard NHS TREs: this review should pay due attention to any duplicated work, duplicated risk, or any identified obstructions to open RAP working or user access.

TRE 29. Manage diverse local datasets by creating and sharing standard data curation tools and methods

As discussed above, local NHS service analysts work with a diverse array of local datasets that can vary widely between regions. Examples include: detailed data about admissions and discharges as bespoke feeds from specific local hospital EHR systems; regular extracts from local authorities about payments for individuals' social care, or more detailed but bespoke social care records from some care providers; intermittent reports from single local services about activity and costs. This presents a profound challenge for any efforts to bluntly standardise all local NHS and care data to a single model. Nonetheless many different data centres, widely separated by geography, will have similar underlying datasets, where shared data management and analysis approaches can be valuable; and nearly all will share common over-arching analytic goals. As per the chapters on Open Working and Data Curation, the best approach to making this curation and analysis activity open, efficient and generalisable is to adopt RAP, and create a small range of standard data management functions and libraries that can operate in any NHS data centre, so that curation work can be shared alongside appropriate technical documentation.

TRE 30. Ensure all local implementations of national or commonly used datasets such as SUS/HES conform to a single standard

As discussed in Data Curation, much local work is done using national datasets such as HES/SUS and GP data. At present the same data is needlessly held in different structures, models, formats, and services in each setting. The National NHS TRE technical team should identify the best flexible approach to harmonisation and this should be adopted wherever local representations of national datasets are required (for example in linkage to local datasets that are not available in a national NHS TRE).

TRE 31. Ensure all datasets extracted from national datasets in NHS Digital are requested using standard data management code

As discussed in Data Curation, much of the data used by local NHS service analysts in local NHS data centres has been provided from NHS Digital; often this is provided in different forms at different times; often this is the product of a discussion, rather than simply submitting the formal specification of a requested dataset in code. This adds to needless duplication of work at all sites and should be addressed by NHS Digital developing a service to accept complex derived dataset requests in code.

TRE 32. Work towards federated analytics with standard local TREs

"Federated analytics" is a secure approach to executing data analysis against locally held datasets in situ, without extracting all local data to a central repository. At its best, a single set of instructions for data management and analysis are written in one location, then sent out to each smaller data centre, where they execute successfully, producing a local version of the results from the analysis using the local data in each setting; these non-disclosive aggregate outputs are then sent back to a central location for aggregation. Successfully executing federated analysis requires that a number of conditions are met: all data in each local data centre must be in a common data model, or capable of being curated into a common model for the purposes of the single analysis, ideally using curation code written centrally; each local data centre must be capable of receiving instructions, verifying that they can be run, and executing them correctly; and each local data centre must be capable of securely sharing completed summary results to the central location for aggregation. A standard local NHS TRE, and even many services that fall short of this goal, can deliver federated analytics in this manner. This approach could be used, for example, to execute a single analysis describing NHS bed occupancy, by calling local NHS data centres containing such data, even in diverse local formats and implementations; or atlases of variation in activity and clinical outcome using data not present in national datasets. Federated analytics is readily achievable and has already been successfully implemented on NHS data in some form, delivering a range of completed analytic outputs in open code projects such as OpenSAFELY and DataShield (the latter in particular with a large existing user-base).

TRE 33. Ensure local analysts use a national TRE wherever possible

Many local NHS service analytics tasks entail using only a local cut of NHS patient record datasets that are currently held at national scale by NHS Digital (in the case of SUS/HES, which is then sent out to local NHS users); or datasets that will imminently be held by NHS Digital (in the case of GP data). This work should all move to be conducted in the National NHS TRE as soon as possible. This will eradicate duplication of risk and cost, and permit rapid collaborative open development of shared tools and learning.

TRE 34. NHS Trusts and Data access Environments

Many NHS trusts have arrangements for academic and service analysts to use their internal data for a range of projects involving research and service improvement. These projects should be encouraged to adopt the standard NHS TRE service wrapper, and standard NHS TRE technical approaches, with all caveats above around thoughtful structured extension to these standards where they are needed.

TRE 35. Listen carefully to local NHS analysts and TRE managers who describe shortcomings in standard approaches; and address these wherever possible.

None of the above should be taken to be a panacea for all possible data uses; there will always be edge cases. Where these present, they should be shared to the National NHS TRE Technical Team.

TREs for national audits and registries

As above, there is a diverse landscape of different national clinical audit and registry projects, mostly focused on using data for service improvement or monitoring, typically run by organisations outside the NHS, but in close collaboration with the NHS, often in adjacent organisations such as learned societies. These services collect a mixture of routine and bespoke NHS data into various datasets, large and small. The datasets are located in a variety of computational environments, with a range of service wrappers and access arrangements. Often data management, analysis, and visualisation is done largely behind closed doors (excepting the final completed output), reflecting the normal practices of the past. All of these projects have been built and maintained over many years by gifted and committed individuals with positive intentions, often as a labour of love: they are to be admired and praised. However, the current dispersed and diverse arrangements for data management are an accident of history, and far from optimal. Many of those involved in audits and registries are technically minded and passionate about better use of data in healthcare. Many of the datasets are structured as "one row per event" and therefore amenable to hosting in a suitably flexible national TRE.

TRE 36. Use the same TRE approach as above

In terms of technical implementation and service wrapper, these projects present essentially the same challenges as local NHS TREs and academic TREs. They are therefore amenable to all the same interventions as above.

TRE 37. Start with Data Pioneers who can demonstrate computational maturity

The best route would be to commence with a Data Pioneer project, identifying between one and three national audit or registry projects that are willing to be resourced to move to a TRE, ideally a national TRE, or use a standard "recipe" for a local TRE, and embrace RAP and modern open working methods. Selection should be based on those with the highest computational maturity: specifically, those who currently have the largest amount of openly accessible technical documentation; the largest amount of openly shared and adequately documented code; and the team that can demonstrate the deepest skills - or potential - for RAP and computational methods.

TRE 38. Review the current Registry and Audit landscape; work towards wider access and use

More broadly these datasets are likely an under-used resource, and a very dispersed set of projects, of very varying scale: it would be wise to commission a deep dive to describe for each project the data flows, the clinical service improvement and research purpose, where the data is housed, who it is accessible to, and recent outputs. This work should inform future investment and TRE moves.

TRE 39. Work towards audits and registries using national NHS infrastructure, RAP, and TREs

The ultimate destination should be that all such datasets are held in one of the national TREs for NHS data, in particular because these projects often involve using patient data without explicit opt-in consent (in contrast to many bespoke data collections for academic research, where active opt-in patient consent has been sought).

Academic TREs

Academic TREs present a complex challenge, as the work is dispersed, with a wide range of datasets, data structures, tasks, funders, users, norms and working practices, and a strong local powerbase around many single projects. Nonetheless TREs should be regarded as critical national research infrastructure, and they warrant strong strategic direction. Below are a range of proposals around implementation and funding.

Academic TRE implementation

TRE 40. Academics should use NHS data infrastructure to access NHS patient records

All academic work on NHS patient records alone should always be conducted in NHS TREs, and be compliant with RAP and open working methods. Patient data should only be transferred out to other non-NHS data centres when that patient has consented for this to be done (for example in consented clinical trials or research studies). Any proposed exceptions to this should be considered under a prespecified exceptions framework created by the NHS TRE team.

TRE 41. Academic TREs should use standard NHS TRE Service Wrapper and governance

There is a very diverse array of governance and access arrangements across a very wide range of DAEs and TREs delivered or funded through the academic community, including a large number of new projects created only very recently. This duplicates risk, duplicates effort, reduces visibility, risks public trust, and risks reinforcing concerns about monopolies over access. All academic DAEs and TREs should use the standard NHS TRE Service Wrapper and governance arrangements. A limited degree of diversity is justifiable for the smaller subset of older projects where, for example, there may be longstanding differences in the commitments made in consent forms to patients participating in a particular study. However, this is no substantive barrier to harmonisation and any small differences could readily be wrapped within a standard approach on all other matters where there is convergence. Any requirements for substantial deviation from the standard service wrapper should be openly disclosed, alongside the justification; these should be regularly reviewed by the national TRE service wrapper team to identify opportunities to extent the standard governance arrangements.

TRE 42. Academic TREs should use standard NHS TRE and curation approaches where possible

NHS data is complex to manage. Wherever any standard approaches have been created by the NHS for TREs or data curation, these should be used by any academic DAE or TRE that is ingesting complex raw NHS patient data with consent. This will ensure they can share in, and contribute to, any curation or analysis code created in the wider community of NHS data users.

TRE 43. All academic TREs should aim to use shared standard infrastructure

Where possible all academic TREs should share a common, core, scalable underlying compute infrastructure to avoid needless duplication in procurement and provisioning: this should be discussed in close collaboration with those who have experience in this domain, including SERP for health data TREs, and the UKRI Science and Technologies Facilities Council for shared infrastructure more broadly. This approach may also help to further minimise historic risks around perceived access monopolies, and lack of interoperability and shared code.

TRE 44. All academic TREs must support, and should require, RAP and open working

Modern, open, collaborative approaches to computational data science are the norm in other academic fields such as physics, structural genomics, structural biology, and more. Code sharing in the health data and electronic health records research community has fallen substantially behind these other fields. Research with data is done by writing code. The code that underpins scientific research using NHS data must be shared under open licenses for scientific review and efficient re-use as in other sectors. At present many TREs actively obstruct these modern working practices. TREs should be regarded as the main way that researchers in this space can be helped to work in modern open ways by default, by making it easy for them to do so. Supporting and requiring RAP and modern open working practices should be regarded as a core requirement for any academic TRE.

TRE 45. Start with Data Pioneers who can demonstrate computational maturity in research cohorts

There have been many attempts by various organisations to harmonise the data hosting and access arrangements around research cohort datasets. There have also been some complex and labour intensive approaches proposed for “harmonising” these diverse datasets. Overall, the most effective first step would be to identify 2-5 cohorts keen to co-develop TRE working, and to deliver their data curation and analysis work in a standard TRE that supports RAP and computational working. By working in the open, executing data curation through RAP processes with appropriate documentation, and using the same openly accessible TRE arrangements as other NHS resources, data harmonisation and federated analytics become substantially more achievable. Participating cohorts should be selected as the most advanced: specifically those who currently have the largest amount of

openly accessible technical documentation; the largest amount of openly shared and adequately documented code; and the team that can demonstrate the deepest skills - or potential - for RAP and computational methods (which may include, for example, detailed shared data management code written using less than RAP methods in R or Stata, or demonstrating an ability to think computationally, sharing code resources, and abstracting out tasks). This work should build on the work of the Longitudinal Linkage Collaboration as that is a strong current locus of such work. It should be implemented as an open competitive funding call for a minimum two year working cycle, and up to a six month delay from award to work commencement to allow for inward recruitment and onboarding of software developers and data scientists to join the existing team with domain expertise.

Academic TRE funding

TRE 46. All funding for academic work on TREs should pass through a single national organisation

Substantial concerns have been expressed by various senior leaders in various sectors around productivity, coordination, and visibility of funding and outputs for TRE work and related data activity in the academic community. TRE delivery, alongside code and methods, is critical national research infrastructure, at the heart of all ambitions to make better use of NHS patient data for public good. All TRE funding activity should be coordinated by a single organisation, most likely inside either the NHS or UKRI. This organisation should have a clear single line of accountability to government and the NHS, and a specific named Minister. It should be open about all accounts, including details on income and disbursements of funds for individual projects. It should be subject to FOI, as an indicator of the organisation's status and accountability, rather than the specific good of FOIs. All funding for academic work on TREs should pass through this single national organisation. Non-government funders of TRE activity, such as research charities, should be encouraged to participate in this open coordination work, as part of their positive contribution to shared national TRE infrastructure.

TRE 47. All TRE and related funding should be openly disclosed

All academic funding for TRE delivery, and delivery of support code within TREs, should be openly disclosed, with adequate technical detail for the community to see the anticipated work and understand its scale. This should include, for each investment:

- The source of funding (e.g. council, programme);
- The amount of funding;
- The recipient (PI, team, organisation);
- The headline objectives;
- 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 48. There should be follow-up on all TRE projects resourced

This should be regarded as a positive opportunity to share outputs, methods, code, insights and technical documentation for others to review, re-use, or improve; and a chance to share barriers encountered for others. It is to be expected that some projects may not meet their initial objectives: this should be accepted as part of the normal process of building in an uncertain and novel space.

TRE 49. Academic work around TREs should be funded through conventional open competition

There is a perception from some in the community that academic funding for TRE work to date has been closed and non-competitive. To address this perception, it is crucial that the majority of all academic funding for TRE work is open to all, via open competition, in a process whereby all groups and organisations can present ideas and proposals in open competition. This work should also reflect the reality that many of the productive academic contributions to TRE work are likely to be on innovative methods and code, from those with contemporary skills in Research Software Engineering and modern, open, computational approaches to data science with NHS records.

TRE 50. Funders should avoid short-term funding for infrastructure

There is a clear tendency for some academic resource on data infrastructure to be awarded on very short timelines. This may reflect: TRE funding standing outside of conventional competitive funding structures; a lack of strategic coordination; some pressure around short-term funding horizons within funders (although this is overcome for other fields of work). Extreme examples include "sprints", where resource awarded must be spent at very short notice, starting work within weeks, and completing spend within months. Short term funding creates numerous problems. It obstructs recruitment and capacity building, in a space where capacity is one of the biggest barriers to delivery. It prevents the creation of a stable ecosystem or culture around code, methods, and projects. Perhaps most importantly, short-term short-notice funding is a regressive model, in that it preferentially channels resource to established incumbents rather than new entrants: the groups best able to spend large amounts at short notice are large academic groups with a shortfall in income from other competitive grants. This approach to funding systematically excludes newer entrants and more junior researchers, who may have the strong ideas and contemporary computational skills needed for innovative work in and on TREs. Wherever possible funding for TRE work should be the same as other competitive research projects, with 2-5 year project duration, and a six month delay from award date to start date, in order to permit recruitment into the project. For all that there may be urgency now, longstanding shortcomings will be fixed more quickly by taking this approach, than by reinforcing the procurement problems of the past.

TRE 51. Funding for TREs should be separate to funding for single academic analyses

Strong TRE infrastructure can only be delivered in close collaboration with single-subject analysts delivering excellent scientific research papers. However, these are nonetheless two distinct activities. The quality of a TRE project should not be judged by single academic publications on single subjects (except as proof that something has been delivered at all); and the two activities should receive clear delineation in funding and roles. This will help to address the challenge, expressed elsewhere that funding ostensibly earmarked for code, infrastructure, and curation can commonly be diverted into funding academic research projects on single clinical research topics, reflecting the current higher status of the latter activity. An approach with clear delineation of funding, roles and recognition between TREs and single analyses will also help to foster a much-needed, clear, independent community with status around delivery of code, infrastructure, and curation.

TRE 52. All best practice and teams should be identified and augmented

There is a historic tendency for academic groups working on TREs and DAEs to be judged by the extent to which they are able to access data at all (reflecting historic challenges around access and IG); or the number of papers they have published. There is a similar historic tendency for TRE teams to want to hold large volumes of data directly, themselves, in single machines under their own control (reflecting historic norms around how value is judged). Lastly there are good grounds to believe that strong technical work on key tasks such as data curation, secure analytics, and other creative aspects of TRE provision, has been done behind closed doors, without open disclosure or recognition (reflecting historic working practices). Insofar as it is possible to create an approach based more on shared underlying compute infrastructure, and collaborative contribution to open code, there is a risk that strong expertise in these existing closed projects will be overlooked. This should be avoided, if necessary with transfer grants to make prior work more open, accessible, and portable.

TRE 53. An overview of prior investments

All new funding via national funders such as UKRI of TREs and related resources should begin with an appropriately detailed open inventory of all prior investment in this space over the past decade, with the positive intention to learn helpful lessons around optimal delivery. This should focus on, for each previous investment: the source (e.g. council, programme); the amount; the recipient (PI, team, organisation); the headline objectives (with reference to contemporaneous public relations material and project proposals); and the outputs (including a link to any papers, code, technical documentation, or live services produced; and a brief description of positive learnings around TRE delivery). It should be recognised that delivering infrastructure of this kind is challenging and that there may often be unforeseen barriers to delivery; open documentation of these will help inform future delivery.

What to fund

Academic funding for TREs should focus on two core themes: (1) support for shared core TRE infrastructure; (2) development of methods and code for core TRE and data management tasks such as curation, secure analytics, and federated analysis, as per the Open Methods chapter.

TRE 54. Standard, national, shared, core compute infrastructure

There is - with narrow exceptions - no technical or regulatory justification for any cohort, centre, university or other organisation to insist that all research data they hold should sit on their own machine in their own data centre. Cloud infrastructure is the standard for storing and accessing the most highly sensitive data, including hospital records with patients' full name and address. Many academic TREs and registry projects could and should share a common, core, scalable underlying compute infrastructure to avoid needless duplication in procurement and provisioning. One national body - likely UKRI, the NHS, or NIHR - should appoint a core TRE infrastructure lead, resourced to create a team that can evaluate the feasibility of a range of standard generalisable approaches. These should not be on nationally owned machines, but rather reflect a standard range of core implementations that can support TRE work for registry and academic projects from a range of commodity suppliers. As above this should be delivered in collaboration with those with experience of prior work in this space.

TRE 55. TRE infrastructure as code and teams

Academic funding for TRE work should recognise that digital infrastructure to support efficient, secure, reproducible, high quality science requires delivery of code, and teams who know how to work with that code. Funding should be focused on delivery of innovative methods and working open code with adequate documentation, following the principles set out for funders in the chapter on Open Working chapter. The focus should specifically be on methods and code that are portable, and can be used in all TREs, both academic and NHS. Work should demonstrate that it has avoided uninformative duplication or overlap with NHS TRE activity, and strong contribution to the methods and code used for NHS TREs. To maximise the talent pool and the range of ideas proposed it is crucial that access to funding for this kind of work is open to all, and not limited to applicants from a specific set of academic groups or educational institutions.

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 to drive a rich, competitive and collaborative ecosystem of code and methodological approaches for key challenges in health data analysis.

Methodological innovation and code for Automated Data Release from TREs

At present all finished tables and graphs produced in a TRE must be checked manually, twice, to ensure that they do not unintentionally contain any potentially disclosive information about an individual. There is a clear user need for automated approaches to this task, and substantial prior art in this space. A programme of work on this topic might focus on questions such as: what are the core abstracted components of the disclosivity checking task conducted by people; which can be automated; what is the state of the best prior art in this space, for example the more mathematically driven work on uniqueness and disclosivity; what theoretical elements can be implemented swiftly; what are the achievements and problems when this is implemented in practice; how can workflow optimise use of humans where they are needed; and so on. This work combines theoretical work on disclosure and privacy engineering; deep domain knowledge around clinical records, TRE design and user journeys, information governance requirements, and epidemiological or NHS service analytics; and pure software engineering skills.

Methodological innovation and code for Data Curation

As in the chapter on Data Curation, there is a wealth of work to be done on the best methods for evaluating NHS data, converting it into analysis-ready datasets, systematic approach to validating EHR data at scale, and so on. This work combines theoretical work on informatics; deep domain knowledge around clinical records, TRE design and user journeys, EHR system design, and epidemiological or NHS service analytics; and pure software engineering skills.

Methodological innovation and code for data minimisation

Minimisation is a commonly used strategy to protect patients' privacy, but those in decision-making roles at data provider organisations have little formal or specific guidance to help them adjudicate on 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, drawing on theoretical work for disclosure and privacy engineering; deep domain knowledge around clinical records; information governance requirements; and more.

Methodological innovation and code for detection of data misuse

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. Applied methodological work and code tools in this space would meet this need, drawing in many of the domains and skills listed above.

Methodological innovation and code to detect unwarranted variation in care

NHS service analysts commonly set out to monitor service activity and clinical outcomes in different organisations, in order to identify which services have the greatest opportunity to improve the quality, safety and cost effectiveness of care, or which services have the greatest to show their neighbours about high quality efficient delivery. There is extensive prior art in this space, and numerous challenges such as avoid over-inclusive or insufficiently sensitive algorithms. As per the chapter on NHS service analytics, there is 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

Federated analytics is a complex set of tasks. Developing new and effective methods requires deep technical skills around research software engineering, but also very deep technical domain knowledge around the kinds of data to be accessed and curated, and the kinds of analyses to be conducted. For example, simple descriptive statistics can be combined from multiple data centres with simple arithmetic; whereas approaches to combining intermediate outputs from complex statistical models require substantial and creative statistical thought around issues such as mixed effects or fixed effects meta-analysis, the best approaches to combining different intermediate elements, and so on; this becomes substantially more complex in turn when moving from a single analysis to a generalisable framework for multiple similar analyses. These are precisely the kinds of complex methodological challenges that must be overcome to deliver federated analytics for complex analyses on complex datasets: they cannot be met by closed working in siloes.

TRE 56. Exceptions to TRE usage

Following the announcement that the GP data extract will only be available in a TRE, some groups began to request exceptions from this rule. Two substantive exceptions to TRE usage are reasonable to consider.

Consented cohorts

Where participants have already given a very large amount of disclosive and private personal information to a given research project, such as a birth cohort; and they have given their consent for their NHS records to be extracted, transmitted, and matched onto their research records inside whatever data management service the researchers are currently running; then it is unreasonable for their NHS records to be withheld from the researchers.

Clinical Trials

Patients participating in a randomised controlled trial have already given written and informed consented for their data to be collected as part of the study: again, this should be respected. There are many opportunities to deliver trial follow-up, and cohort follow-up, inside a national TRE: but there is no need for this to be mandated, as patients' preferences and consent should be respected.

Where there is a clear operational requirement to disseminate data outside of a TRE for purposes that will improve patient care, without consent, it may be reasonable to do so - under a robust exceptions framework - after all relevant IG and ethical processes have been followed, where it can be shown that additional steps have been taken to preserve patient privacy, for example by robustly minimising the data, and releasing data for a sub-sample of the total population. It is important for context to note that census data - for example - is not shared in this way; and that TREs are achievable and bring many benefits. A robust exceptions framework can gather examples of where dissemination was deemed necessary, and review annually the need for any extension or modification to the available TREs to ensure that such dissemination is minimised in the future.

TRE 57. Address TREs for Artificial Intelligence, but as a separate workstream, funded by existing AI resource

TREs with the core features described above will readily support all analysis using traditional analytical or epidemiological research techniques. Analysis and research involving techniques that fall under the heading of Artificial Intelligence - particularly unsupervised machine learning - present some different technical challenges for TRE design that need to be considered. These challenges primarily relate to compute power; the use of specific tools such as GPUs; and export controls, because exported random forest models can (in ways that are hard to detect) sometimes contain disclosive patient data. Overcoming these challenges is essential if patient privacy is to be maintained while using NHS records for AI research at scale.

It is crucial that any bespoke work for an AI TRE is resourced separately from the core TRE work that meets the current analytic needs for the NHS service analytics and research community. AI has very substantial potential (some of it already well demonstrated) around imaging data; it has some potential around EHR data. However, AI work is attention-grabbing, and can sometimes crowd out other work. Substantial national resource has already been devoted to work on AI in healthcare, much of it committed to tasks other than an AI TRE for NHS data. There is a strong case for some resource being spent on creating core infrastructure that allows AI teams - both public and private - to innovate on NHS data in a secure TRE sandbox. This will release economic gains in due course, as with other TRE work. However, the crucial core strategic challenge for better use of NHS data is the overdue need to create strong foundational infrastructure for conventional data analysis in TREs to support current analytic needs with non-AI methods. Overcoming these challenges will likely help deliver an AI TRE.

If separate resource can be found to develop an AI TRE technical team, they should: rapidly evaluate current TRE offers and how well they manage disclosiveness when releasing AI-based models, with open delivery of their detailed technical findings; validate and reproduce prior research evaluating disclosiveness to ensure the results are accurate; expand the specifications of current offerings to overcome any limitations discovered; and implement a TRE capable of safely supporting AI. This work should be delivered through the same methods set out for conventional analytic TREs: approaching the project as methodological innovation and open code with technical documentation, rather than closed black box services; open competitive funding to find the best talent; and leadership from those with appropriate skills and proven delivery on data infrastructure platforms, rather than an excessive focus on single academic skills.

Objections to TREs

While TREs bring many benefits around privacy, modern working methods, and efficiency, some concerns were expressed to the Review team by the community about this way of working. These fell broadly into three categories. Firstly, there were concerns about delivery of a TRE. Secondly, there were a range of specific concerns about specific use cases. Thirdly, and most commonly, there was a general preference from some to continue with the old method of data dissemination to multiple off-site locations, but with larger datasets and more end-users. This latter ambition is relatable, but ultimately

unrealistic. Dissemination has clear disadvantages. Professional groups and campaigners object to it strongly. The care.data programme tried to expand extraction and dissemination without success in 2013. The most recent GP Data for Planning and Research data collection plan was suspended in 2021 and only re-railed after a commitment for TRE-only access. Each of these attempts at extraction and dissemination for GP records resulted in well over a million patients opting out of their data being used. Lastly, a practical mitigation - TREs - is achievable, and the norm in other sectors, including all ONS work on the census. Overall, it is not feasible that better public relations, as sometimes suggested, could address these problems and make expanded extraction and dissemination successful.

Below, to inform future discussions, are the specific objections that have been raised around TREs during the course of the review.

“TREs cannot be used for linking EHR data onto bespoke data collections for research cohorts”

There are no barriers here, with two options: data can be transferred outside of the NHS into any academic setting wherever there is robust patient consent to do so; and national TREs should have the facility to import research data at reasonable scale.

“TRES cannot be used for randomised trials”

There are no barriers here. For trial data analysis, participant data can be exported where there is robust informed participant consent to do so; and randomisation schedules can be imported into TREs for analysis within the TRE. Similarly TREs can be used to identify patients for possible participation in a randomised trial: feasibility counts can be done on national or regional data that is pseudonymised; then, where there is a desire to identify individual patients to be approached (either directly or via their healthcare team) this can be facilitated by a simple re-identification service that can work back from pseudonym to real identifier, where the user has permission to do so.

“What about situations where you need to re-identify a patient for a clinical risk issue”

Where necessary this can readily be done in a TRE, with a re-identification service.

“AI is hard in a TRE, because it requires extensive bespoke hardware”

There are various models for a TRE tailored to the specific needs of AI: helpfully AI is a field that has seen very substantial recent investment whcih can readily support the deployment of such a service to make NHS patient records available for AI work while also preserving patients’ privacy.

“Some regulators say they want to export a model from an AI TRE to validate it on a different dataset”

Random forest models (under consideration here) when exported can compromise patient privacy as they can contain - often in forms hard to identify - elements of patient records. However, this is unlikely to be a common issue. If a model has been developed against the entire England population dataset, then it is unclear what different external data MHRA would be able to access to validate it on. If the model was developed on a sub-sample of the England population in an AI TRE containing the whole population’s data, then it can be validated within the same TRE on another sub-sample. Lastly, while there are strong ambitions around AI regulation, MHRA do not appear to do this external validation work regularly, if at all, at present.

“Regulators, and industry, want to link together data from lots of territories”

There are very few other countries with any capability to offer GP or other EHR data on the scale of England; it is hard to conceive of many countries agreeing to the wholesale bulk export of large volumes of their citizens’ health records; and there are less disclosive options around federated analysis for most if not all use-cases. Nonetheless it would be useful to evaluate any specific edge cases for this concern.

“TREs cannot be used for local NHS data”

Local NHS data presents some additional challenges around the diversity of the datasets, but there is a clear path forwards for addressing security, consistency, and duplication in a stepwise fashion as discussed elsewhere in this chapter.

“TREs cannot be used for multimodal data and linkage outside a single TRE”

The use case here is, for example, getting EHR data in an environment where it can be analysed alongside the 40 petabytes of genomics data in Genomics England. Where an organisation has consent to extract patient data, it is reasonable for EHR data to flow there. More generally, a more appropriate paradigm is likely to be that data is minimised at source in one TRE, and the minimally disclosive transfer is subsequently made between TREs: so an analysis using sparse genomic data, but detailed EHR data, might be done better in an EHR TRE than a genomic TRE.

“It is unlikely that international regulators will find it acceptable to carry out assessments within a UK TRE”

There is no reason for anyone not to work in a flexible and secure analytic environment that meets their analytic needs.

“TREs are hard to use”

It is crucial to ensure that TREs meet real user needs; but also to recognise that some forms of deeper and more flexible interaction with data will require deeper technical skills on the part of users; it is also important to recognise that some of the challenges reported by users working in TREs are caused by the scale of the data accessible in a TRE, rather than the fact of working in a TRE.

“TREs are hard to build”

It is clear that there have historically been challenges: for example, at the time of the team’s review in Summer 2021, more than a year after project commencements, the 3 substantial TRE investments led by Health Data Research UK during COVID-19 (ICODA, DECOVID (with the Alan Turing Institute), and the BHF-HDR / CVD-COVID TRE) had produced no research papers or openly accessible code (other than one paper describing the data accessible through NHS Digital, with some accompanying analysis scripts). Nonetheless elsewhere there are also success stories, such as the well-established ONS SRS with its new public health datasets during COVID-19, the excellent work at SAIL/SERP, and very substantial delivery outside of health data research in adjacent communities such as genomics and physics, where open collaborative working with modern computational data science techniques has been the norm for many years. With procurement led by technically skilled teams, clear lines of accountability, clear ambitions, and clear oversight, there is no doubt that health data TREs can be built rapidly by, for, and with the NHS, and made accessible to others in the research community.

Chapter 05

Information Governance, Ethics and Participation

Goldacre Review

Summary

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

Confidence from all can be increased 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.

Rationalise current processes

Alongside this substantial change, there is also a need to address the complexity caused by a diverse range of partially overlapping rules, organisations, and processes that plainly cause a substantial amount of delay and distress to a large number of researchers. In the detailed text below are a range of suggestions around common paperwork, common meetings, and a clear common map of all processes agreed by all relevant organisations in the system. This kind of work is important to help applicants navigate the processes effectively to deliver important analytic work; and to help organisations and individuals have the confidence to share information when it is appropriate to do so. The recently published IG Framework for Integrated Care includes tools and templates to capitalise on good practice, and spread it.

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 (PPIE) 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 following 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.

Background

Over the course of the review - from interviews, focus groups, and desk research - it became clear that the research and analytical community is very, very frustrated with the current Information Governance framework: the combination of laws, regulations, policies, and ethical guidelines governing access to and use of health data. The team heard multiple examples of research with substantial patient benefit being blocked by the complexities, duplications, delays and contradictions of multiple legal, regulatory, professional, and ethical restrictions. Researchers and NHS service analysts can spend months – sometimes even years – trying to get multiple necessary permissions from various parties including trusts, ethics committees, GPs, NHS Digital, the Health Research Authority, individual patients, NHS England, and the Information Commissioner's Office, for even low-risk research projects.

The experience of the LAUNCHES QI Study

LAUNCHES QI, a project initiative in 2018, aimed to link five national data sets to generate understanding about Congenital Heart Disease services, with the intention to: describe patient trajectories through secondary and tertiary care; identify useful metrics for driving quality improvement; and explore variation across services.
In a paper published in the BMJ Open, the researchers describe the process of applying for approvals from the separate NHS data controllers, as well as for permissions from the Health Research Authority, and NHS Digital.

In total 47 documents were required for the data application processes, comprising 384 pages. These were required by 11 data controllers or departments and submitted 162 times in total. Each application form asked for similar study information, but each required different wording, structure, and detail, and some seemed unfit for purpose – designed for clinical trials and other interventional studies involving human participants and ill-adapted to data-only studies. The resulting confusion meant that each data controller asked for alterations and further information between one and nine times before approval was given. This was all in addition to getting research ethics approval from the HRA and gaining Section 251 approval.

Gaining permissions, including university and ethical approval, took 8 months, and the data applications took between 3 and 7 months. Acquiring the data took a further 7-10 months. Once this linkage was completed, the researchers received further funding to start a new – but related - study on the same dataset. The approval process to re-use the dataset took a further 2 years and, at the time the researchers published the case study they were still waiting for approval from NHS Digital.

Given research funding is often time-limited, with timelines agreed before research commences, these kinds of delays can mean that research projects are abandoned – denying the public of the potential benefits.

Taylor JA, Crowe S, Espuny Pujol F, et al. BMJ Open 2021

Researchers, analysts, policymakers alike all recognise the need for strict regulation to meet the goal of protecting patients. EHRs contain the most personal and sensitive information about individuals. As outlined in detail in the Privacy and Security chapter, the privacy threats posed by the use of EHR data for research and analysis are real and should never be underestimated or dismissed. Working with this data is a huge privilege, and the individuals to whom the data relates must always be treated with the utmost respect. This means that access to and use of the data should always be carefully controlled. However, there is an overriding feeling that the level of restriction and caution generated by the "spaghetti junction" of regulations is disproportionate and overly burdensome.

"Our experience suggests that IG processes are a ubiquitous challenge in delivering timely and high-quality research with the potential to make a positive impact on health. Currently, such processes are often lengthy, circuitous, opaque, and inconsistent, in a way that creates duplication and wastage of time and effort and is not proportionate to the nature or degree of risk involved."

- Interviewee

The current system

Information Governance (IG) is often unfairly regarded as an obstructive or bland discipline, but in reality it is a complex multidisciplinary project requiring skills in analytics, IT, ethics and IG. At its best there is a clarity of purpose and an energetic embrace of role and accountability, with IG professionals working with others to leverage maximum benefit from information, enhance patient care and improve services while protecting patients and remaining compliant with the law.

One key challenge is the complexity of the current governance framework, with multiple partially overlapping remits and processes, and no clear strategic oversight, plan, or coordination between each element. There is a strong sense - from interviewees, and from our own experience of speaking with various actors in the various governance processes - that individuals administering specific aspects of the system have deep expertise within their own component, but often do not recognise that those outside of their specific organisation or role may find it confusing; and often do not have a clear understanding of how their components relate to, or overlap with, the adjacent components that analysts will also have to address.

The use of personal data collected in other sectors is largely governed by simpler set of rules - with their own challenges nonetheless - set out in the UK General Data Protection Regulation and the UK Data Protection Act 2018, overseen by the Information Commissioner's Office (ICO) and the Department for Data, Digital, Culture, Media and Sport.

By contrast the collection, storage and use of health data is governed by a multi-layered set of overlapping, duplicative and sometimes contradictory policies, regulations, and ethical guidelines managed by the Department of Health and Social Care, the National Data Guardian, the ICO, the Health Research Authority, Medicines and Healthcare Products Regulatory Authority (MHRA), NHS Digital, NHS England, the NHS Transformation Directorate, local trusts, individual universities, GPs, hospitals, and a range of other bodies granted powers by the preceding organisations.

Relevant legislation includes the Health Service (Control of Patient Information) Regulations 2002, NHS Act 2006 (specifically section 251), the Health and Social Care Act 2012, the Data Protection Act 2018, the UK GDPR (General Data Protection Regulation), Medical Device Regulations, and the Human Rights Act, as well as the Common Law Duty of Confidentiality. Relevant national policies include the Caldicott principles, the five data sharing principles from the NHS Transformation Directorate's Centre for Improving Data Collaboration, the UK Policy Framework for Health and Social Care Research, the national data opt-out, and ethics guidelines set out by the Health Research Authority (HRA). Finally, there are hyper-local policies set out by individual data controllers – of which there are thousands, including each individual General Practice. In addition to this there is a strong role for personal judgement by the individuals responsible for the implementation of these diverse rules, regulations and guidelines in diverse organisations, as the rules themselves are sometimes loose and open to interpretation.

This complexity likely reflects a range of causes, including a longer deeper history of large volumes of data being used in healthcare; the fact that health data is often very detailed, very disclosive, and very private in its nature; and the fact that there are a very diverse range of organisations and stakeholders throughout the complex ecosystem of the NHS. There are currently various initiatives to address some of the complexity, including the Health and Social Care IG Panel, the IG Portal and the Red Tape Challenge, each aiming to provide consistency on policy, guidance and advice, with the aim of ensuring data is accessible and that participants in the system are aware of their duty to share, alongside other pressures.

This is very welcome. Combined, the various overlapping rules and frameworks create a patchwork of activity and outcomes that can often be inconsistent: some organisations act (arguably) recklessly; some are extremely cautious and obstruct necessary work; and sometimes specific patients' informed, consented, documented and very clearly expressed wish that their NHS records should be shared with a specific research team is actively obstructed.

This layering of multiple interacting organisations, laws, regulations, and polices makes it almost impossible for analysts, patients, or those employed to protect patients to see the wood for the trees. It makes it hard to see what the single obstruction is, for any single project, or field of work. It is barely possible for any one person, group, or organisation to have complete oversight of the combined governance framework, its performance, whether it is achieving its objectives, and whether it is being consistently and proportionately applied. Instead, each individual, group, or organisation focuses - for relatable reasons - largely on the policy over which they have complete control, with little consideration for how this might interact with another policy or legal requirements. Consequently, researchers frequently find themselves bounced between different organisations and data controllers for long periods of time, without ever hearing a clear yes or no. As one senior representative of the HRA said:

"Even if i or other experienced researchers know how one bit of the forest works, navigating the complexity of the multiple regulators and sources of approval is what befuddles researchers (and the public)."

- Interviewee

The consequent delays to project initiation are a deep source of frustration for researchers and can be the cause of significant opportunity cost to the healthcare system, as well as actual wasted cost to funders, by preventing the continuation of important, and sometimes even urgent, healthcare research. The team were given multiple, detailed, credible stories from individual researchers of projects that had been delayed for years, including several where projects were entirely abandoned, as they had only received approvals long after the funding and medium-term employment of relevant staff on the work had finished.

These problems are well established. Researchers have been sharing their concerns for close to twenty years. It would be wrong, therefore, to suggest that nothing has been done to try to improve the situation.

"There is evidence that UK health research activities are being seriously undermined by an overly complex regulatory and governance environment....New regulatory bodies and checks have been introduced with good intentions, but the sum effect is a fragmented process characterised by multiple layers of bureaucracy, uncertainty in the interpretation of individual legislation and guidance, a lack of trust within the system, and duplication and overlap in responsibilities."

- Quoted in Harmon and Chen (2012) from A New Pathway for the Regulation and Governance of Health Research (AMS, 2011)

The draft NHS data strategy includes commitments to simplify information governance, building on the work started by the Information Governance Portal, and to introduce legislation which will create a statutory duty for organisations within the health and care system to share anonymous data for the benefit of the system as a whole. Additionally, the NHS White Paper states that the Government is currently considering a range of actions including making changes to NHS Digital's legal framework to introduce a duty on NHS Digital to have regard to the benefit to the health and social care system of sharing data that it holds when exercising its functions and clarify the purposes for which it can use data. In addition, the ICO has been developing guidance to provide greater clarity on the various research

provisions included in existing legislation; and the Department of Data, Digital, Culture, Media and Sport is currently consulting on a number of proposed legislative changes which would bring together research-specific provisions and ease the burdens placed on researchers, including health data researchers. Whilst it is beyond the scope of this review to analyse these proposed changes in detail, the underlying intent should be commended and all those with interest and expertise in this area should be encouraged to engage with the relevant government consultations.

In short, this is a profoundly challenging space where many organisations have tried to improve matters, by changing the rules, over many years. However, the team were also told on numerous occasions by researchers and stakeholders of all levels that it is not necessarily the design of individual policies, regulations and ethical guidelines that are the problem, but rather the way in which they are interpreted and implemented by the organisations and individuals responsible for making decisions allegedly in accordance with governance rules.

A culture of caution

Alongside the complexities, contradictions, and overlap of the various different individual regulatory frameworks, many who engaged with the review also felt that rules - which typically require substantial personal interpretation - could often then be applied with excessive caution.

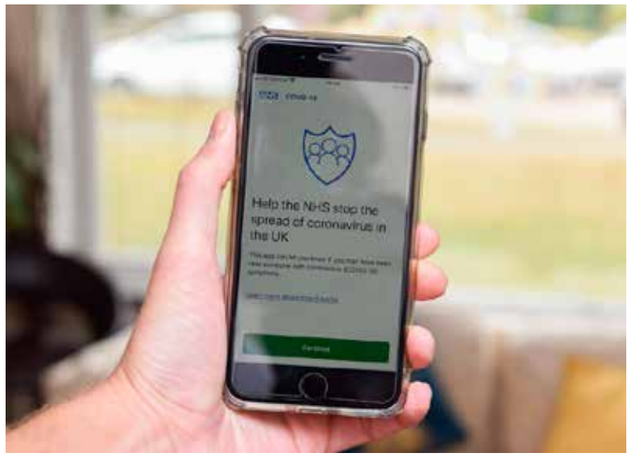
"Sometimes people don't understand the rules so it's easier for them to say 'the data protection law won't allow me to share data with you.' Usually, the data protection law would actually say you can share as long as you do xyz."

- Interviewee

"Lots of people will say the GDPR is a blocker. It wasn't supposed to be. It was supposed to be about sharing data. The blockers are all interpretations. Some of it is there are too many independent bodies."

- Interviewee

The unintentional negative impact of this prevailing culture of caution – which Allen and others refer to as "privacy protectionism" – has been thrown into stark relief by some examples from the period of the coronavirus pandemic. Regulation 3 of the Health Service (Control of Patient Information) Regulations 2002 allows confidential patient information to be processed in various new ways in relation to communicable disease and other threats to public health. It does this by providing the Secretary of State for Health and Care with the legal power to require certain organisations to process CPI for purposes related to communicable diseases. On 20th March 2020, the Secretary of State used these powers to issue NHS Digital, NHS England & Improvement, all healthcare organisations, arms-length bodies, local authorities and GPs with notices requiring them to process CPI for the purposes of COVID-19. These notices - colloquially known as "the COPI notice" - have been essential in enabling several life-saving research projects, such as work by the national chest imaging database, and UK Biobank. They have now been in place for 18 months and, although not permanent, have demonstrated that the system can be permissive when it needs to be. Yet data access and governance problems have persisted. Several researchers reported incidences of well-funded COVID-19 data science projects being blocked from accessing data, even when covered by the COPI notices. As one senior researcher said:



"It is mindbogglingly difficult to understand why there is a culture of withholding information out of fear of sharing and breaching confidentiality when there is a public health crisis and the information that is being shared is incredibly low risk."

- Interviewee

Instances of data being withheld, even when there is a clear legal basis, reinforce the idea among researchers that the barriers they hit when accessing data are not just regulatory, but also cultural or organisational. This leaves researchers and analysts feeling beleaguered, with the sense that they are presumed to be doing something illegitimate, or with bad intentions; and forcing them to spend much of their time negotiating and completing paperwork, rather than using their data science skills.

"There is an assumption of maleficence that everyone who wants access to data has a nefarious intent."

- Interviewee

"It feels to me being an analyst in the system rather than being a leader, it feels like we spend our whole lives dealing with red tape rather than actually dealing with data. We are just fighting the system constantly when we should be just getting on with doing actual work."

- Interviewee

From interviews, desk research, and consideration of the current arrangements around sharing data, it seems that this caution flows from three sources: an incorrect assumption that the public are against data access for research; anxiety caused by indeterminate rules requiring individuals to take responsibility for personal judgement calls; and a historic lack of safe mechanisms to securely share disclosive patient data.

Anxiety caused by a belief that patients are against data sharing

The team was unable to find any formal research on regulators' perceptions of public preferences, but were repeatedly told by data users that they felt those responsible for interpreting the governance framework seemed to assume that patients and the public were in general opposed to researchers and analysts accessing data, other than in exceptional circumstances. There has been a wide variety of research on public preferences around data sharing. Overall, the range of socially acceptable uses for health data is broad, provided certain conditions are met in relation to security, transparency, and accountability (see box).

Summary of current research on patient and public attitudes towards secondary use of EHR data

The organisation Understanding Patient Data maintains a library of current research into UK public attitudes towards the use of health and patient data for secondary uses. As of this review, the library covers the period September 2018 to August 2021 and lists the following relevant publications:

Putting Good into Practice: A public dialogue on making public benefit assessments when using health and care data, published by the National Data Guardian in April 2021.

Public deliberation in the use of health and care data, the findings from the OneLondon citizens' summit which took place in March 2020

Foundations of Fairness: where next for NHS health data partnerships? a report commissioned by Understanding Patient Data into public opinion regarding third-party access to NHS data, published in March 2020

Patients' and Public Views and Attitudes towards the Sharing of Health Data for Research: A Narrative Review of the Empirical Evidence', published in the Journal of Medical Ethics by the authors in November 2019

Public views on sharing anonymised patient-level data where there is a mixed public and private benefit, a report published by the Health Research Authority in September 2019

Giving Something Back": A Systematic Review and Ethical Enquiry into Public Views on the Use of Patient Data for Research in the United Kingdom and the Republic of Ireland', published by the authors on Wellcome's Open Research site in January 2019

Who benefits and how? Public expectations of public benefits from data-intensive health research, published by the authors in Big Data & Society in December 2018

Investigating the Extent to Which Patients Should Control Access to Patient Records for Research: A Deliberative Process Using Citizens' Juries published by the authors in the Journal of Medical Internet Research in March 2018

The findings of these various reports and research projects, across many different groups of stakeholders, are remarkably consistent. They show that patients and publics are generally supportive of EHR data being used for legitimate health research purposes, including research designed to improve clinical services. However, this willingness is not unconditional, it depends largely on 'trust' which itself is dependent on a number of factors, including: who is conducting the research; whether those individuals are regarded as having legitimate motivations; and whether security measures around data access are adequate.

The NHS, universities, and charities are perceived as being the most trustworthy organisations, as they are perceived to have the most legitimate reasons for wanting to conduct research. In contrast, commercial companies are seen as being the least trustworthy, as their primary motivation for conducting research is perceived to be 'profit.' However, use of health data for research by commercial companies - including pharmaceutical companies - is not seen as a red line, provided this is done transparently, and clearly for patient benefit. Transparency and a clear public benefit are not only important in the

context of commercial research: they are two of the most commonly cited elements of 'trustworthiness' across all studies.

"Benefit" is generally a broad category for patients and publics, and has been shown to refer to a wide range of outcomes, such as an improvement in access to services, or developing new drugs for long-term conditions. Even secondary benefits, such as secondary economic effects, are recognised as being legitimate reasons for conducting research. What matters more than the specifically defined benefit, is its translation from research into reality. Patients and publics currently feel as though this translation process is being hampered by politics and bureaucracy, and are keen to have greater visibility of the 'translation process.'

Other than privacy, transparency, and public benefit, commonly cited conditions for 'trustworthiness' include: data security, control, information, responsibility, accountability, and fairness. Several of these - but not all - can be immediately and robustly addressed by the use of TREs, as shown by the results from a recent citizens jury examining public attitudes towards various data sharing initiatives set up during the pandemic (see separate box).

While there is much in this research that should reassure regulators around public preferences, there is also much for researchers and analysts themselves to act on, for example:

- Ensuring that their work has clear benefits
- Ensuring that these benefits are clearly communicated
- Ensuring that their work is transparent, reporting activity, methods and results
- Ensuring that their work is executed in a secure manner.

Anxiety caused by a lack of determinacy in Information Governance rules

Many researchers and analysts who had interacted with those administering the rules felt that these individuals were sometimes anxious, on a personal level, about the decisions with which they were tasked. While there may be room to improve individual aspects of individual rules, there will always be an inherent challenge in creating fixed rules that can perfectly account for all possible complex scenarios. As a consequence, individuals may feel exposed, as they are required to make personal judgement calls on complex and important issues involving substantial risk. They will naturally feel that they carry personal responsibility for the decisions they make. These anxieties will only be worsened, in turn, by the fact that projects can comply perfectly with the rules, but nonetheless sometimes go on to become extremely controversial. The care.data programme is one example of this: a large project with a robust legal justification for collecting, processing, storing, and analysing data, which nonetheless failed to earn public trust and was ultimately cancelled, resulting in more than a million people opting out of their data being used outside of their GP practice.

The care.data programme

Care.data was a substantial programme of work involving NHS patient records data announced by NHS England in 2013. The intention was for NHS Digital (then called the HSCIC) to extract GP data, pseudonymise it, link it with hospital data such as HES, and then make this data available for researchers for the purpose of improving patient care. The programme had a robust legal basis in the Health and Social Care Act 2012 and was legally compliant with data protection law of the time. This sound legal basis was not sufficient to garner public and professional support. There was a substantial backlash against the programme, and after multiple delays it was shut down in 2016.

Care.data is frequently cited as a cautionary tale about use of data, and a situation to avoid repeating. Retrospectively, it has been asserted that the problems were principally caused by a failure to mount an adequate communications campaign. One widely recognised example involves a leaflet sent to all households in England announcing the programme: many people (including several journalists) complained that they did not receive this leaflet; some apparently assumed it was junk mail and disregarded it. When people did access the leaflet, they found it did not contain the words 'care.data', so it was unclear how it related to the programme of work being critically discussed in the media.

Alongside communications were a range of other challenges. The programme was on a scale that had not been seen before, providing access to detailed GP records covering the full medical history of the whole population: this represented a substantially more detailed set of records than had ever been previously shared. In addition, the programme was launched and communicated before there were clear plans in place around who could access the data, on what basis, and for what purposes. This meant that researchers hoping to champion the work could only state that the NHS should be trusted, rather than point to a new institution or set of principles that would guide access and provide assurance. Concerns were expressed regarding privacy invasion, and use of the data by commercial or insurance companies, which were augmented when journalists examined other NHS work with data. They found that datasets such as HES had apparently been disseminated for uses that were not well known at the time, and fell outside the range of uses that some commentators and patients found acceptable. As these were entering the public domain for the first time, without substantial prior discussion, this exacerbated general concerns around all data access and dissemination.

Overall, care.data shows the importance of clear communication, but also the need for clear rules around access, good mitigation for privacy risks, transparency around access, and due recognition of when a new programme is at a scale that pushes the limits the current "social license".

Anxiety caused by the mechanisms used for data access

When data is accessed inside a TRE, it is easy to monitor what is done with it, to ensure that any analyses fall within the permissions granted, to prevent onward dissemination of patient data, to ensure that only permitted individuals have access, to obstruct any invasion of patient privacy, and to swiftly detect any attempted misuse. At present, most current data access is by dissemination of pseudonymised records for use at off-site locations. As discussed in the chapter on Security and Privacy, this approach presents avoidable risks around privacy and security (which can be mitigated by greater use of TREs).

When using dissemination, much of the risk management is up-front, at the point of decision-making around whether to disseminate or not. A substantial degree of faith must be put in contracts, and the promises made by organisations and individuals around appropriate use of data after they import it: while audits can be done (and are sometimes done) of the sites where data is received, these should not be given undue weight. Some involved in decision-making around dissemination stated that - because of the need to trust the party receiving the data - a large component of their decision making was a judgement call about whether the recipient was trustworthy. This judgement call, they said, would rely on factors such as whether the recipient had much to lose from misuse, and whether they were already well known and respected in the wider ecosystem as users of data. These judgement calls speak, again, to the extent to which the rules are only a component of the decision-making process around dissemination. They are also likely to obstruct new entrants: in the past, users of EHR data tended to be only a small number who had worked on it for many years; now, and in the future, there is a clear ambition (and need) to broaden the pool of data users to include new teams and organisations, for example those with strong general data science skills, applying new methods and tools.

In short, data dissemination requires decision-makers to trust the recipients of data, and is inherently more risky than TRE access. Because of this, it makes decision makers substantially more cautious about data access than they might be for more secure mechanisms around data access. Trust, as a security and privacy technique, presents challenges when there is a requirement for scale.

The impact of TREs on public trust

TREs present a simple solution to many of the problems outlined. Rather than relying on trust, contracts and promises, TREs facilitate more robust proof of security and privacy: they allow all data use to be monitored, ensuring that all analyses are within the users' permissions; they prevent onward dissemination of patient data, to ensure that only permitted individuals have access; they can obstruct invasion of patient privacy; and they can swiftly detect any attempted misuse. Strong TREs also provide a mechanism whereby detailed logs of all activity can be disclosed for external scrutiny, providing a robust, credible and public account of all users, all projects, and their implementation.

By providing a more secure mechanism for data access, TREs can help decision-makers feel more confident about permitting users to access data. Alongside the material fact of TREs providing greater privacy safeguards, there are also good grounds to believe that these are understood and recognised by the public, for example in a recent citizens' jury that examined data-sharing lessons learned during the pandemic (see box).

Citizens' Juries on data sharing in a pandemic

In 2020 the NIHR commissioned a set of three online citizens' juries about health data sharing in a pandemic, in collaboration with NHSx and the National Data Guardian for Health and Social Care. These took place between March and May 2021, with each of three juries spending eight days listening to detailed evidence and deliberating on three national data sharing initiatives which were introduced to tackle the pandemic: OpenSAFELY; Summary Care Record additional information initiative; NHS COVID-19 Data Store and Data Platform.

These juries were commissioned to address key policy questions arising from the ways in which health and social care data sharing changed during the pandemic and from new initiatives developed to facilitate sharing. Questions included: should these data sharing initiatives, created under temporary legal powers to tackle COVID-19, continue beyond the pandemic; if so, then for how long; and who should make the relevant policy decisions? To answer these questions, each jury watched a presentation from the expert witnesses for each initiative, and were then given the opportunity to ask questions of each witness. After the presentations and Q&A sessions from all the expert witnesses, jurors were asked the following questions about each of the initiatives:

How supportive are you of the decision to introduce this data sharing initiative in 2020 as part of tackling the COVID-19 outbreak?

What are the most important reasons to be supportive?

What are the most important reasons to oppose the initiative?

For how long should the initiative continue?

By whom should these decisions be made?

How could or should the initiative and its uses be usefully changed in the future (if at all?)

What actions, if any, could be taken to engender greater public trust in the initiative?

What are your reasons for opposing the initiative?

OpenSAFELY, the only initiative constructed principally as an open TRE, was by far the most strongly supported by all three citizens' juries: 100% of jurors were supportive of the decision for it to be introduced (77% very much in support, 23% broadly supportive); all three had more jurors "very much in support' of the TRE than for any other initiative; and 87% of jurors believed that it should continue for as long as it is useful, provided the decision to keep it running is made by an independent advisory group of experts and lay people.

When asked why they supported the OpenSAFELY TRE more strongly jurors explained that they considered it to be more transparent, more secure, less risky, and therefore more trustworthy. This is demonstrated by these example reasons given for supporting it: "The initiative does not transfer or store data, meaning we do not have another platform holding vast quantities of data and the accompanying risk of it being leaked"; "[it] protects against misuse of the retrieved data via multi-level access, audit trails, publishing of code and no direct downloading or

accessing of the data and publishing outputs"; "It is a software platform that doesn't require the moving or downloading of data, so data cannot be edited or copied, and researchers do not need to access the data in order to analyse it, ensuring confidentiality and minimising usage of sensitive information and maximising safety and security." These comments demonstrate that the public understand the technical design of a strong TRE, and appreciate the positive impact this has on trustworthiness.

Managing anxiety and risk with TREs

The complex current system of IG and other permissions - which is slow, and materially obstructing good work to improve patient care - has evolved over time to manage the inherent security and privacy challenges of data dissemination. It is not reasonable to apply all aspects of the data dissemination governance process to TRE access. TREs very substantially mitigate many of the risks around data access, and have been well understood by patients, policymakers, professionals and the public as doing so. In the detailed recommendations below are a number proposals that may help to simplify the IG framework created for data dissemination, as many have tried to do before. More important, however, is the proposal to create a "two track" approvals process, with TRE approvals granted under a more efficient IG process created to proportionately manage the substantially lower risks when patient records are accessed through a strong TRE. This should not be taken to imply a "free for all" around data access: for example, it is possible to conduct a piece of research that is securely executed, but nonetheless also unethical, or harmful to patients. Appropriate controls must therefore still be maintained around purpose, and ethics.

Other barriers to data access

Outside of IG and ethics frameworks, a range of other access barriers were cited, as discussed below.

Monopolies, resource, and recognition

Several senior and junior researchers and analysts gave credible examples of situations where they had asked an organisation for access to a given raw dataset, to conduct a particular analysis, and been told that this was not possible on grounds of information governance; but then subsequently saw, many months later, that the same analysis had been conducted, by the same organisation that had refused them permission, apparently having given their own analysts permission to do the same work. In their view, this represented a conflict of interest (COI) and a misuse of complex information governance rules by individuals and organisations who had other reasons for wanting to prevent access to data they had either collected, or currently had some form of control over.

"Monopolies over data are one of the largest hurdles that must be overcome... "Guarding" of data, be it for reasons of commercial advantage, academic competitiveness or resistance to the idea of commercial involvement, is harmful to collaboration and only adds expense or prevents research outright. This hampers the effective use of data and stifles innovation (by both public and private sector organisations)."

- Interviewee

The team was also told by individuals working within organisations that they had no doubt that internal staff sometimes obstructed access to data by external users, because they felt that the final analysis and published report was the output regarded by the wider community as having the highest value, and that collecting and curating data was lower status. Lastly it is clear that some academics regard aggregating a large amount of data as one route to stable resource for their other academic work in a complex competitive space.

"It feels like there is a bit of a club at the moment in terms of getting access to data. there are friendly faces. Sometimes people who are known get preferential treatment it feels like. There are also worries about what if you find things that are critical of the NHS or the system etc. We are an elbow in the side sometimes."

- Interviewee

This is a challenging issue that has received little prior discussion: it should be recognised as a sensitive topic, that may be regarded by some as controversial; but nonetheless requires resolution. Patients' best interests are best served by having the best analytics, from a diverse range of skilled teams, on all data.

It is important, first of all, to recognise why an organisation may want to retain some degree of monopoly over analytic outputs. Data collection and management is complex, skilled, and costly work that currently has unhelpfully low status as an independent activity: this problem, and resolutions to it by funders and anothers, has already been discussed in previous sections on Open Working, Data Curation, and TREs. It was clear from discussion that some in the system incorrectly regard the marginal cost of sharing data as being close to zero, once it has been created. This would certainly be the case for someone placing a collection of existing PowerPoint files online, or on a USB drive, for a colleague to review and re-use. Complex EHR and research data, however, is very different.

The costs of collection alone can be very high. There are strong examples from the Cohort Studies community (see TRE chapter) where teams have been resourced to collect and then share data outside their group; but this is still not a universal norm among funders. Once raw data has been collected, the costs of data management and adequate technical documentation are similarly high. This problem is often compounded, in turn, by the fact that these are tasks are often also invisible under current working practices: they tend to be bundled unhelpfully within the process of delivering a single analytic output, within a single team, commonly with little systematic sharing of code or technical documentation, even internally (see Open Working and Data Curation). This working style means that any internal team working on an internal dataset may conclude that no outside user could work appropriately with their data, or that it would cost them a lot of support time to help those external users. They may be right to have these concerns; albeit that the problems are caused, or augmented, by technical documentation on datasets being historically prioritised too lowly.

"If it's patients' and the collections are done through public money, then these collections need to belong to the public. The collections cannot belong to the doctor or researcher. This applies to registries too."

- Interviewee

This raises a number of serious issues around incentives, and appropriate use of regulatory structures. These issues are incorporated into the detailed recommendations for reform of the information governance and access arrangements at the end of this chapter. However, the appropriate general principles should be as follows:

- It is inappropriate for information governance processes to be used to obstruct data access for other reasons;
- People who have invested time and effort on collecting or managing data that is widely used should be able to access resource to make their work for all sustainable;
- Data collection and curation should be regarded as independent skilled activities with status on a par with writing final data reports;

- The marginal additional costs on an organisation when sharing data should be priced appropriately, and passed on appropriately.

Overall, the risk of monopolies and conflict of interest around data access should be openly recognised and discussed, and directly managed by:

- Ensuring those granting permission for access (for example on an organisation's data access request panel) are independent, or include a range of independent external users who are aware of the issue of COI;
- Including a space on organisations' data access request forms prompting the data controller or processor to declare any COI they may have
- Providing a facility for appeals in situations where COI may have obstructed access;
- Conducting research to understand what incentivises sharing of data.

It should be noted that TREs, which substantially reduce the risks around data sharing, may make it less feasible for organisations to use IG as a barrier to access; however, this will only bring forward the need to address the other underlying barriers to data access discussed.

Anxiety about performance management, and public accountability

A related form of conflict of interest arises around healthcare organisations who are concerned that the patient records data they hold may be used "against" them, for example to execute data analysis projects that are intended to deliver performance management metrics. This is another complex and sensitive issue that has not received substantial open discussion, but has nonetheless driven a range of choices and practices around data access.

There is no doubt that some GPs have been concerned that the planned General Practice Data for Planning and Response data collection may be used to monitor their practices' activity and outcomes, or to raise criticisms of individual clinicians; and that this has slowed progress on the programme, alongside the concerns about patients' privacy when their pseudonymised records are disseminated off-site to multiple locations. Concerns were also expressed by government and other analysts interested in health, but outside of direct NHS employment, that their access to NHS data was limited because some in the NHS were reluctant for them to be able to independently examine activity and outcomes. There was also more specifically a concern that data may be used to monitor activity in a haphazard of inaccurate way, which may require time-consuming rebuttals, or have adverse consequences on clinical work: there are indeed recent historic examples of poor public communication around patient records data being used for performance management nationally.

This has resulted in several uneasy "halfway house" arrangements over time. It has contributed to the culture of closed working practices around NHS service analytics, as the assumption of secrecy around some specific performance measures for individual healthcare organisations has bled over into an assumption of secrecy around the very methods and code by which such metrics are conceived and calculated. This in turn has undoubtedly limited the quality of such metrics, by limiting the individuals and teams who can contribute to their creation and constructive critical review. A range of detailed recommendations to resolve this issue are made at the end of the chapter.

Concern about commercial users

In media and public discourse around access to NHS data, two principle concerns dominate: appropriate safeguarding of patients privacy; and the notion that NHS data is being "sold" for commercial use. TREs help to very robustly address the former concern: they make it possible for data to be made accessible to a broader range of innovative users from the public and private sector, with appropriate safeguards, to a much greater extent than the previous paradigm of data dissemination. For this reason, along with many others, TREs should be energetically created and maintained by the NHS.

However, TREs do not address the second concern, which is essentially around the ethics and practicalities of commercial use. This is clearly a complex and controversial area. It is also run through with misconceptions. For example, sometimes NHS organisations seek cost recovery to cover some of the work involved in making data accessible to a commercial or semi-commercial user. The sums involved for this transaction are typically very modest. It is therefore unhelpful to see this cost-recovery styled by some as NHS data being "sold". Indeed, many expressed the view that the sums taken for cost-recovery reflect the NHS under-charging, and failing to take an appropriate stake in the benefits arising from use of data collected - at great expense - by the NHS.

Similarly, there is a tendency by some to view commercial users of data as being uniquely less trustworthy, uniquely prone to conflicts of interest, and uniquely likely to conduct misleading analyses: this is similarly unhelpful. There are certainly legitimate concerns around commercial conflict of interest: the chair of this review has written extensively on the subject in the past; but also on the misuse of data by a range of other actors including academics. As context for this, peer-reviewed research from 2018, 2020, and 2021 (led by the chair and published in the BMJ and The Lancet) shows that clinical trial results are commonly left unreported, in breach of widely accepted ethical guidance and legislative obligations; but that pharmaceutical companies now perform very substantially better than their academic counterparts, and are substantially more likely to correctly comply with their trial reporting obligations, in data from trial registries from both Europe and the US.

Furthermore, there appears to be a widespread lack of understanding about the uses to which commercial companies put NHS data: these may include marketing, which should be considered separately, but also include a range of tasks that plainly deliver substantial public benefit. For example, when side effects are spontaneously reported for a given treatment, regulators will typically approach the relevant pharmaceutical company and require them to conduct pharmacoepidemiological research, using complex statistical models in electronic health records data - often from the NHS, using subsets of the GP data - to evaluate the extent to which a given adverse outcome is more or less common in recipients of different comparable drugs, or with comparable medical histories. This is plainly a positive use of NHS data by commercial companies.

Where there are any concerns about companies with a commercial interest conducting flawed analyses of NHS data, TREs again provide a very useful protection. There are two principal concerns with misleading analyses by actors with a COI. The first is that studies can be flawed and biased by design, in deep technical ways that are often hard to discern: a TRE that shares all code executed as a public log acts as a substantial protection against this, and a clear source of audit to facilitate its detection. The second is that analysts sometimes engage in "p-hacking", running multiple slightly different analyses until they get a preferred answer: again, any analyst from any sector attempting this in a TRE that shares open logs will find that they are immediately detected in doing so.

Lastly, the most under-discussed misapprehension in this space is likely to be that the NHS can leverage substantial revenue from simple direct sales of NHS data to commercial vendors for single purposes. Overall, discussants from the pharmaceutical industry, academia, and life sciences agreed that the marginal cost savings for them of single acts of transactional access for NHS data - such as follow up in a clinical trial alone, or a pharmacoepidemiological

study - were likely to be modest, making the prospects of national income from such activity similarly modest.

By contrast to the prospect of any single sales, by far the largest overall economic benefit for the NHS and the nation is likely to come from whole systemic packages: not clinical trial follow-up data alone, but rather platforms that can execute the entire pathway of a clinical trial in an efficient, digital manner, with proportionate governance. Similarly, as the everyday processes of the NHS come to be increasingly digitised, and there is a gradual move towards better harmonisation of data (as outlined in the chapter of Data Curation, among others) then new prospects for innovation open up, some of which may be transferable to other settings and nations. In this regard, the greatest commercial benefit is likely to come from the whole system: well-curated data, in accessible and performant TRE platforms, with appropriate technical documentation, alongside a digitally competent NHS, with clear entry points that are technical as well as "negotiations and meetings", and above all a workforce with deep competencies that combine generalist data science or software development skills with deep domain knowledge on health data, its strengths and weaknesses, its provenance, and its effective use in analytics and innovation. Indeed, representatives from industry expressed very deep scepticism about the value of current national investments aiming to make NHS data more accessible to them, which appeared to make data access contingent on collaboration and meetings with academic gatekeepers that they felt - in many cases - they neither wanted nor needed.

There is a very extensive array of detailed policy work with large teams in diverse settings across government around the best mechanisms for appropriate cost recovery, and commercial engagement; these arrangements sit outside the terms of reference of this review. Brief recommendations are given at the end of this chapter.

Exclusive commercial arrangements

The team were told by analysts of situations where, in their view, access to data from an NHS organisation had been unreasonably withheld because that organisation had an exclusive commercial relationship with another organisation around access to patient records data. At the same time, some of those involved in such arrangements expressed concern that that the financial sums involved, where they believed they knew them, often seemed to be comparatively trivial; that the income for the single NHS organisation did not outweigh the network disadvantages of these apparently or genuinely exclusive relationships; and that the limited revenue also did not outweigh the bad feeling caused within the relevant NHS organisation, and other organisations, by such arrangements. It was also stated that the true extent of exclusivity or revenue was often unclear, and could be the topic of rumour rather than straightforward disclosure. Overall exclusive relationships are likely to be disadvantageous for the NHS; this is discussed in the recommendations.

Multiple data controllers

The NHS is a complex network of dispersed individual organisations, each of which has complex contractual, practical, and historic relationships with other organisations locally, regionally, and nationally. As a consequence, those wishing to conduct research or analytics often find themselves needing to seek ethics approval, IG approval, or contractual agreement, from a very wide range of organisations including individual Trusts, and individual GP practices.

This was expressed by many researchers and analysts as a profound source of concern, completely obstructing many projects, and making others prohibitively expensive. The team was given examples of projects where many thousands of patients had given permission for their GP records to be accessed for research, but that the researchers then had to negotiate separately with each GP practice to get their own separate approval for records to be released. UK Biobank is a cohort study with hundreds of thousands of participants who are making themselves available for scans, blood tests, genome sequencing, questionnaires, and so on. All of these participants have given written informed consent for their GP records to be accessed by UK Biobank, and linked to the other study data, for access by researchers globally. Despite this consent, UK Biobank's team has been required to negotiate separately with about 6,500 GP practices – since each is the data controller for some of the participants – in order to obtain separate permission from each GP for each participant. UK Biobank is therefore still not able to make GP data available to support research into the causes and prevention of many different diseases. Recent special access to these data under pandemic legislation, solely for the purposes of COVID-related research, has now demonstrated just how valuable GP data is for studying the determinants of disease; and, therefore, how much is being lost by the continuing failure to make these data available more widely.

Requirements such as this create a range of problems. Firstly, they create a very substantial administrative burden for researchers who are delivering work for patient benefit. Secondly, they create substantial administrative burden for clinicians in GP practices who are required to spend time reviewing and considering detailed and complex documentation on data sharing for a small number of patients, when their time is valuable and required for direct care. Lastly, they risk arbitrary decisions being made, that may not reflect the best interests of patients, practices, the NHS, the public or any other party, in a hurry, or without complete information, or without complete background knowledge, or without the due consideration that a complex data sharing agreement truly warrants.

This in turn has led to a range of problematic outcomes. For example, there is now a complex patchwork of different GP research data extractions, each taking large volumes of NHS GP electronic health records data out of hundreds, or sometimes thousands of GP practices. GPs give consent for this data extraction by setting a flag in their electronic health records system; patients are not asked for consent. Typically, these detailed electronic health records are pseudonymised by the removal of direct identifiers, and then made available for download by large numbers of researchers in various different countries and organisations.

Setting aside the fact that this does not reflect emergent ideal practice around using a TRE to protect patients' privacy, there are important questions to be addressed around whether this is the best way to manage access, when considered in the practical context of how approvals are granted. For example, there are numerous different organisations running these datasets; some GP practices have agreed to all of them extracting their patients' data; some have agreed to none; some have agreed to some, but not all. Are some GPs really making the decision to participate in some, but not all such projects? If so, on what basis are they distinguishing between them? Similarly, the decision to permit all patients' records to be downloaded off-site for use by a range of external analysts and researchers is a consequential one: are all GPs who have not agreed to this making an informed choice to withhold their patients' records from researchers, after reading all the relevant legal documentation and arguments, or have some not yet considered the issue? Have all those who have agreed to their patients' data being extracted definitely read, digested, and separately considered all the relevant legal documentation around this data sharing? And do they have the detailed contextual knowledge necessary to separately assess them all?

Overall, it is hard to argue that it is a sensible or proportionate use of NHS GP time - or even practice manager time - for 6,500 individual GP practices to separately consider all these issues and legal documents, on multiple occasions, for multiple different research projects, both general purpose and singular studies. This requirement also imposes substantial workload and risk on individual clinicians; but it is a natural consequence of the current legal reality that each GP practice is the Data Controller for their own patients' records, and must separately grant permissions. Addressing this problem was one ambition for the national General Practice Data for Planning and Research data collection, by creating a single focus for subsequent access requests. Now that this dataset has been made "TRE only", it has very strong prospects of progressing, and may alleviate some of this pressure; additional approaches to reducing the burden on clinicians are proposed in the recommendations below. Similar issues arise for separate permissioning from Trusts for their own patients' data.

Patient & Public Involvement and Engagement

Exploration of PPIE as a topic in itself was not a specific request in the terms of reference for this review, nor does the team claim to be experts in this domain. Designing, conducting, and analysing the results of high quality PPIE is a specialist skill in and of itself. However: health data represents people; each EHR used in each analysis represents an individual person; and each individual data point – a diagnostic code, referral, prescription record or similar – represents a moment in a person's life that may have had deep meaning for them at that time, or a continued impact on their experience of life. It is, therefore, absolutely essential that these individuals are respected, and that their autonomy is protected when health data is being used for research and analysis.

This is why PPIE is vital, and why it is at the core of all work on data access, data analysis, and all related areas. Well-designed, meaningful PPIE can help to ensure that patient and public trust in research is maintained, and that the individuals to whom records relate are treated with respect and dignity; co-designed PPIE, and co-designed research projects, can also improve the quality of research. Patients and public representatives are the experts of what it is like to experience the care of the NHS, to live with specific conditions, or to care for loved ones experiencing ill health. This means that they often know better than any independent researcher or analyst the most important research questions, the right outcomes to measure, and the best way to ensure that the outputs of any and all research delivers on its ultimate goal: patient and public benefit.

For these reasons and others the most useful, successful, and impactful health data research projects are often those that design with, and for, patients and the public from the very beginning; that involve a diverse range of representatives in every decision, from what data to request, to how to interpret results and disseminate findings; that listen to and act on the advice, feedback, and input of these representatives; and that treat their values, beliefs and experiences as crucial to success as well curated data, performant software, well executed code, or a carefully designed statistical model. More fundamentally, by holding all researchers and analysts across the health data research ecosystem to this standard of PPIE, the NHS can ensure that its treatment of patients' data, just as its treatment of the patients themselves, is in accordance with the principles and values set out in its constitution, especially that:

- The patient will be at the heart of everything the NHS does
- The NHS is accountable to the public, communities and patients that it serves
- Everyone counts

It is therefore essential that the importance of PPIE is never underestimated or dismissed, and that it is never seen as secondary to the data analysis itself. This is why it is positive to see research funders, ethics committees, and data access committees placing such significant emphasis on the importance and value of PPIE. To understand what exactly these various bodies should be looking for when assessing PPIE plans, the team engaged with expert patient and public representatives and leaders of the PPIE community throughout the review through interviews, focus groups, and written submissions. The team also reviewed the literature on the topic and studied examples of PPIE highlighted to us as being best-in-class. Reflecting on the insights gained from this process, and building on the Consensus Statement on Public Involvement and Engagement with Data-Intensive Health Research (see box below), there is a clear need to champion PPIE that is:

Participatory

Involving open questions and engaging activities designed to elicit a deep understanding of participants' values, beliefs, and lived experiences.

Example participatory PPIE activities include:

- Co-design workshops
- Role-play using user personas to develop user stories
- Empathy mapping

Inclusive

Accessible to all, with appropriate accommodations made for those with additional support needs and involving participants who reflect the diversity (in every sense of the word) of the population that the NHS serves.

The types of considerations that might be needed include:

- How best to recruit a diverse range of participants, especially participants from marginalised groups
- How to make sure all activities are device-agnostic (and don't require participants to have access to a specific piece of technology)
- Which venue to use for in-person activities that is suitable for all accessibility requirements
- Whether slides or other visual or reading material can be adapted for those who use screen-readers or other accessibility tech

Discursive and Deliberative

Involving the exchange of information, opportunities for learning, and respectful debate. Held in a safe space where all opinions can be voiced without fear of judgment, and where changing ones opinion is accepted and encouraged.

Example discursive and deliberative PPIE activities include:

- Citizens juries
- Semi-structured focus groups
- Online discussion forums

Meaningful

Predicated on the idea that the advice, opinions, and views of the participants will be acted upon and used to guide the research – even when this involves significant change to the direction of travel.

Ways to ensure this is happening include:

- Giving participants the opportunity to prioritise research questions
- Giving participants the opportunity to comment on study protocols, governance arrangements, and findings

- Feeding back to participants how they influenced the research
- Ensuring participants are acknowledged in published papers and, where appropriate or desired by the individuals, included as co-authors.

Recurring

Cognisant of and responsive to the fact that opinions, beliefs, and values are not static in either time nor space, but vary significantly depending on context and changes in experience.

Methods to keep participants engaged throughout lengthier research projects include:

- Involving participants in regular project meetings
- Maintaining a public-facing website that is regularly updated
- Scheduling engagement activities to be repeated at regular interviews
- Providing a means of continuing asynchronous conversations with participants, for example: email, WhatsApp, or other messenger services

Consensus Statement on Public Involvement and Engagement with Data-Intensive Health Research

In 2019 a deliberative process involving an international group of stakeholders and experts in patient & public involvement and engagement, resulted in the publication of the Consensus Statement on Public Involvement and Engagement with Data-Intensive Health Research. The statement sets out 8 key principles to establish the importance of PPIE in data-intensive research and ensure it is practiced in a consistently high-quality way. The key underlying premise is that the public should not be characterised as a problem to be overcome but a key part of the solution to establish socially beneficial data-intensive health research for all.

According to the statement, PPIE should:

- Have institutional buy-in
- Have clarity of purpose
- Be transparent
- Involve two-way communication
- Be inclusive and accessible to broad publics
- Be ongoing
- Be designed to produce impact
- Be evaluated

Whilst abiding by these guiding principles will not always be possible, they provide a baseline expectation that high-quality engagement should be deep, discursive, and deliberative.

Aitken, Mhairi, Mary P Tully, Carol Porteous, Simon Denegri, Sarah Cunningham-Burley, Natalie Banner, Corri Black, et al. 'Consensus Statement on Public Involvement and Engagement with Data-Intensive Health Research'. International Journal of Population Data Science 4, no. 1 (12 February 2019)

Fortunately, the need to conduct PPIE that fits these criteria is well-recognised by the health research community at large, and there are multiple examples of excellence for researchers to learn from and be inspired by. These examples include, but are not limited to:

- The Genomics England participation panel
- The OneLondon citizens' summit held to inform the OneLondon local integrated health and care record

- The citizens juries commissioned by NIHR Applied Research Collaboration – Greater Manchester – to explore health data sharing during the COVID-19 pandemic.

These examples all reflect engagement on different topics and different types of data; all involved different modes of engagement. However, they all have a common underpinning commitment to deep, open, and reflexive conversation designed to ensure patients and public representatives feel like the designers of the health data research ecosystem rather than simply its beneficiaries. The more researchers, analysts, and others using health data for purposes beyond direct care can be encouraged and supported to conduct PPIE of this nature and calibre, the more patients and the public will trust in health data research and, consequently, the more everyone will benefit from its outputs. Some analysts and researchers expressed concern that they were sometimes asked to do extensive PPIE prior to receiving funding, or without support, and that this made engaging positively with the process of PPIE more challenging. Some specific suggestions on these topics are given below.

Recommendations

A range of detailed recommendations are given below across the following themes:

- Enhanced usability for IG and ethics processes
- Two-track approval for TREs
- Regulation and legislation
- Addressing specific roles and uses
- Patient and Public Involvement and Engagement

Enhanced usability for IG and ethics processes

It is crucial that the various diverse IG and ethics frameworks protect patients and the public. While maintaining high standards there are various simple changes that could help to make the system clearer, faster, and more navigable for applicants.

IG 1. Create a single form for all ethics, IG, and other data access permissions

Researchers are regularly required to describe the same aspects of the same project in multiple different ways for multiple different organisations to approve multiple different aspects of their permissions separately. Researchers should be able to fill in a single form from a single starting point: the relevant sections of the form on different aspects of IG, ethics, and related issues should be accessible to each of the different relevant organisations from whom approval is sought. The current patchwork approach is duplicative, inefficient and risks issues falling through the gaps.

IG 2. Streamline the number of approvals meetings

Researchers regularly have applications rejected because of the need for clarification of another aspect of their permissions from another approval body. As with the single form above, where possible, whenever one project requires multiple permissions from multiple bodies, this should be addressed at a single meeting where all relevant bodies can collectively review and discuss all aspects of applications in one go. This will be more or less practical depending on the regional coverage of each committee: however, coordinating timings, and overlap, should be readily achievable in many settings. Different organisations can and should take responsibility for different considerations within their own remit: but there should be open conversation between them, and researchers should not have to repeat themselves, or experience delays because of complex concerns about overlap or non-overlap between different varieties of decision-making body.

IG 3. Get researchers in the room

Researchers regularly have applications rejected because of a misunderstanding, or a need for a clarification; when this happens, they may have to wait several months for the chance to have their application re-considered, after addressing the misunderstanding. Whenever a meeting is taking place, the applicants should be ideally be present: this should pose no obstruction to open and considered deliberation. Failing this, the applicant should be informed of the time and date that their application is being considered, and the committee given a telephone number where they can call the applicant in case any factual aspect of their application requires disambiguation.

IG 4. Create an arbitrator for disagreements over specific access requests

Disagreements over access should reduce when data controllership has been streamlined across the NHS, and where strong TREs reduce privacy risks. However, issues may still arise, especially when trying to link to datasets outside of NHS control. In these circumstances, an arbitrator should be able to step in and make the final decision. This, in combination with other recommendations, should help tackle issues related to conflicts of interest and monopolistic behaviour among those holding patient data.

IG 5. Create a single map of all approvals

The approvals process may seem simple to those administering it: even then, they may only have clear oversight of their own component. For those navigating the system, it is often profoundly confusing and complicated. A single map should be created of all required approvals, with links to access further detail at each stage.

This should be reviewed by the organisations implementing each regulatory step: they should collaborate on and contribute to the descriptions and roles of their own processes. This map should be regularly reviewed for accuracy, and for opportunities to simplify the system.

IG 6. Provide rapid unambiguous guidance when approval is not required

When a piece of work is believed to sit outside the remit of a given organisation or process, the organisation responsible for that aspect of governance should confidently provide a clear statement that their involvement is not required. Complex aspects of governance may sometimes be a matter of judgement calls; organisations administering the governance processes are in a strong position to make those judgement calls and share their insights. (For clarity, this is often seen already: it should be welcomed, applauded, and its value recognised).

IG 7. Establish two modest Centres for Regulatory Science

Regulation of research, and information governance, will always be complex. The trade-offs inherent in different options will always be challenging to unpick. It is unrealistic to expect that this will ever go away, and it is unhelpful that there is very little "commons of knowledge," advice, or critical review of current rules in public. At present, there is largely only "folk knowledge" on the part of applicants, rather than a rich commons of knowledge. In the US, the FDA (Food & Drug Administration) commissions, resources, and tasks a small number of practical research units evaluating the efficacy and utility of regulations, legislation and processes managing risks in healthcare, including through the Advancing Regulatory Science Initiative, and the bioethics research communities at universities. These have challenged and improved regulations in several areas. We should replicate this effort on a modest scale: UKRI or NIHR could fund small groups responsible for producing detailed critical reviews of interesting cases for discussion; critical analysis of existing and proposed regulations; clear advice for individual researchers and organisations seeking access to data; and feedback to the policy community on what is and is not working. The objective should be to create a rich, practical, public, critical commons of knowledge around governance. This should be modelled on the welcome expansion over the preceding decades of professional medical ethicists in universities. Staff should be multidisciplinary, technologists, clinicians, researchers, social scientists, and lawyers with experience of regulation and information governance.

IG 8. Establish a clinic to help users who are blocked on data access

MRC has previously funded excellent work at their Regulatory Support Centre to give support and guidance to individual MRC-funded groups encountering barriers to their work around data access or usage. UKRI and/or NIHR should expand this work, or augment a group from the Centres for Regulatory Science, to create an open problem-solving unit that invites reports on blocked projects, locally and nationally. This group should be focused on helping patients, clinicians, commissioners, and researchers overcome the practical, technical, regulatory, and cultural barriers they encounter when they are trying to access data, with practical guidance on how to 'unblock' access in ways that are safe, legally compliant, ethically viable and technically feasible for each situation. The aim should be to share this work, and create a growing library of themed insights and solutions, for each project describing the barriers encountered, and how they were overcome. This group should provide an annual open report to policymakers describing their work and any systemic issues or recurring themes that they have encountered.

Two-track approval for TREs

TREs substantially reduce the privacy risks - but not the ethical risks - involved in using NHS patient data. This should be reflected in the governance arrangements around access to data through a TRE.

IG 9. Create a 2-track approval system to incentivise use of TREs

The current complex IG arrangements were principally devised to manage the privacy risks that are inherent to more insecure methods of data analysis through dissemination, where there is a greater risk that data could be leaked, illegally viewed, or otherwise misused. TREs substantially address many of these issues; and many governance questions can be addressed once in a formal review of the TRE, rather than bespoke for each individual project. The NHS Transformation Directorate should conduct a formal review of which existing safeguards and processes can be accelerated or retired for projects conducted exclusively within a strong TRE. This should result in a substantially less onerous and faster access and approvals process than for work using conventional and less secure methods with data dissemination. This is proportionate, will explode productivity, and will actively incentivise better, safer ways of working.

IG 10. Maintain excellent standards around governance issues not addressed by TREs

Not all aspects of governance are rendered obsolete, or less important, by the introduction of TREs. Regardless of how data is analysed, the purpose for which it is used, and potential harms from this, must still be subject to careful scrutiny for each single analysis. Ethical review and PPIE should therefore persist for single analyses, subject to the efficiency improvements described above. However, any ethics and PPIE work on the facts and processes of data access itself, rather than the single specific analysis, should be done once only for the TRE as a whole, wherever this is feasible and appropriate for the single analysis in question.

IG 11. Review the National Data Opt Out Policy after TREs are established

The National Data Opt Out policy was introduced in response to the crisis in public trust caused by the problematic implementation of care.data in 2013. It has, however, been problematic in implementation, inconsistently applied, and can give patients the impression that their data will never be used for purposes other than direct care. It should be reviewed, but only after a strong national TRE has been established for use of GP data and other commonly used national datasets, as above. If patient data is only ever stored securely, never directly 'seen' by researchers, and used transparently, then there may be fewer circumstances in which it is logical to allow people to opt out; or opt-outs could be reviewed to cover different classes of use, rather than different classes of data flow. Any changes should be carefully considered with meaningful input from patient and public representatives; following adequate research into the motivations of opt-out usage; and retrospective changes to current opt-outs should be handled with great caution. Nonetheless, TREs present an opportunity, if not a guaranteed route, to carefully develop a new accord with the public around restrictions on use of their data.

IG 12. Uphold the commitment that the NHS Digital GPDPR dataset will not be disseminated

This dataset is unprecedented in detail and coverage, it cannot be meaningfully pseudonymised by the removal of direct identifiers. It is very welcome that this data will now only be accessible in a TRE. It is critical that this commitment is adhered to.

Regulation and Legislation

IG 13. Revise the definitions of "anonymous" "identifiable" and "linked" data; add a new category of "pseudonymised but re-identifiable"

There has been a widespread misapprehension across the system around pseudonymised data, and an excessive confidence in the privacy protections provided by the removal of direct identifiers. This misunderstanding has been pivotal in a range of problematic decisions around risk management. However, it is in part a consequence of the currently available formal categories of risk for datasets relating to patients. There have historically been attempts to create simple nomenclatures that describe whether data is identifiable or not on the basis of the data alone. Current commonly used categories of data are, in brief, limited to: "anonymous data" (for example, "3,200 people died of cancer last year in Norfolk"); "identifiable data" (for example, data with name and address openly stated for each patient); and "linked data" (where one dataset has been linked to another).

These three categories are insufficient to describe the challenges faced in secure management of detailed NHS electronic health records data, and in particular they do not capture one of the most commonly encountered forms of data. There is a fourth category - "pseudonymised but readily re-identifiable data" - which should be formally added into common parlance, regulation, and legislation. The extent to which this data presents a privacy risk is a function of the data itself, and the context in which it is being accessed: if disseminated off-site, where a user can interact with it as they please, with no logs of their activity, this pseudonymised rich data is profoundly insecure; in a robust TRE with barriers to viewing disclosive data and logs of all activity, then it is securely held and presents few privacy concerns. It is crucial that the system recognises and describes this category of data as a central privacy risk to be mitigated: recognising this will allow the system to make informed choices and earn the trust of campaigners, professionals, and the public. This issue is increasingly discussed, and is the subject of a current ICO document on anonymisation, put out for consultation in Autumn 2021. However, the concept remains overall rather nameless in regulation and legislation, falling between stools, with work largely guided by interpretations and guidance: it should be a central focus of legislation and regulations that aim to address privacy issues.

IG 14. Consider including health data in the Digital Economy Act

Increasingly health data needs to be linked with other administrative datasets to understand, for example, the social determinants of health. This would be made easier if all public datasets were governed in the same way. There will need to be extra protections for health data, as outlined in detail elsewhere, but it may be unnecessary to have an entirely separate legal framework. The Digital Economy Act (2017), and the inclusion of health data within it, should be formally evaluated with this in mind. As with the above recommendation about the National Data Opt Out Policy, any changes should be carefully considered with meaningful input from patient and public representatives. Health data, linked to non-health data, should only be accessible in a robust TRE that prevents direct access to patient data and shares informative logs with the public to ensure complete transparency about all uses to which the data are put. ONS has deep technical knowledge and history in this space, and should be regarded as a beacon for future work, including through their forthcoming Health Strategy. As with so many recent projects where data access has been facilitated in a time-limited fashion by pandemic legislation, the Public Health Data Asset work between ONS, the NHS and Public Health England shows the power of wider linkage.

IG 15. Appropriately sanction those who are caught deliberately and maliciously attempting to re-identify individuals in patient records

There are numerous accounts of people inappropriately accessing fully identifiable records without consent or legal basis, or otherwise misusing patient records. When this happens - in the case of, for example, celebrities being admitted into hospital - the individuals are often caught, and disciplinary action is taken. However, while it continues to be possible to download detailed, pseudonymised but readily re-identifiable patient records to local machines the true scale of inappropriate use is unknown. Wider use of strong TREs will make it easier to detect deliberate attempts to reidentify individuals, but this may not be a sufficient deterrent if the consequences remain minimal. Regulators across the system, including professional bodies, the ICO, MHRA (Medicines Healthcare Regulatory Agency), and HRA (Health Research Authority) should coordinate to develop and implement appropriate fines and sanctions for those caught deliberately breaching patient privacy. A strong deterrent for any individual person misusing health data must be regarded as a crucial component of any robust regulatory framework if it is to achieve its objectives. This should not be regarded as a barrier to wider and better use of data to improve patient care: it should be regarded as a key facilitator of that objective.

IG 16. Disclose all data flows leaving NHS organisations in one place

Throughout this review we have received detailed descriptions of many substantial bulk flows of NHS patient data for service research and research, including complete patient records, outside of local NHS and associated organisations, including GP practices, on the basis of approval by local organisations. Current recipients are diverse and include NHS organisations (such as NHS Clinical Commissioning Groups); NHS adjacent organisations such as GP Federations, NHS Commissioning Support Units, Academic Health Sciences Networks; large and small scale commercial providers of analytic assistance to NHS organisations (individually or as groups) running their own data centres with NHS patient records; and a broad range of less visible participants including private providers of research services, private providers of case-finding services, private providers of administrative services to GP practices, and several privately and publicly owned GP research datasets, extracting and granting onward access to millions of patients' data without patient consent. From our interviews, it was clear that many in the system are unaware of the extent of this data dissemination. This may be because attention has historically focused more on national data moves administered by national organisations such as NHS Digital. Some of these data flows were in our view, and others', disproportionate to the apparent stated objectives of the work. There are strong grounds to believe that the work done or intended to be done from these data flows could be achieved more effectively, and more securely, in a national secure TRE. The NHS Transformation Directorate and DHSC should commission or conduct a detailed open review of these data flows to establish in some detail whether they are safe, proportionate, and can be replaced with more secure options. This work may raise some currently undocumented concerns; however, these are likely overall to form the basis of a stronger case for a single GP data flow into a national secure TRE. Once reviewed, all data flows should be logged in a central system that is visible and public.

IG 17. Create a central repository of DPIAs, DSAs and related documents for local NHS data flows

Many of the NHS data flows currently in place, especially for local projects, are not well known even to clinicians, policymakers and researchers. There are detailed governance requirements around paperwork and disclosure for individual organisations, however information about these flows is often not easily discoverable. The NHS Transformation Directorate should establish a central indexable repository of DPIAs, DSAs, privacy notices, and other related documentation for all small local data flows so that this information can be searched and viewed by interested parties. This should not impose any additional burden on NHS organisations, as it entails solely the sharing of existing documents. Alongside greater transparency and visibility it is likely that this will also help build a more robust commons of knowledge around best practice for such activities, and their appropriate documentation.

IG 18. Produce boiler-plate templates for patient consent for data use and dissemination

There will be certain circumstances where it may be necessary to provide local downloads of patient data, for example, for patient follow-up in clinical trials. In these circumstances the patients must have given explicit consent for their data to be accessed for research purposes outside of a TRE. Current mechanisms for gaining such consent are variable in quality. Central provision of clear boiler-plate templates for consent to disseminate patient records will raise standards and improve public trust.

IG 19. Simplify the rules governing use of posthumous data

Posthumous medical records are an essential resource for almost all health data research and analysis: studying the records of people who have died is one of the most effective ways to understand how to prevent death. Yet the rules governing the use of posthumous records are confusing and inconsistent. Different types of record are kept for differing lengths of time, in different circumstances, with different access mechanisms, before being destroyed. One team should be charged within he NHS Transformation Directorate or similar to harmonise all

requirements, with a view to improving research; data should be held securely in TREs as with living patients' data; there should be an assumption that records are preserved, with a clear commitment that they will be securely managed for all time, to the same standards as for living people; patient opt-outs should be respected for this class of data to the same standards as those used for living people.

IG 20. Address the "multiple permissions" problem

NHS patient data is a vital and powerful resource for improving the quality, safety, and efficiency of healthcare. This requires that data is accessible. The current requirement to obtain permission separately from each organisation for each act of data sharing is a substantial practical barrier to better harmonisation, and better access. It is driven by the current legal reality that each organisation is the Data Controller for the records they hold. Two options may help to make this more manageable, subject to detailed legal and policy evaluation, and public and professional consultation. Firstly, consideration should be given to whether a national organisation could become Data Controller for a copy of all NHS patients' records, to be held only ever in a secure national TRE (as is planned for some GP records in the General Practice Data for Planning and Response programme), where it can be worked on for the purposes of service improvement, academic research, and foundational work into data curation and harmonisation. Secondly, and less effectively, consideration should be given to creating an "approvals pool", where large numbers of Trusts, GP practices and other NHS organisations can voluntarily nominate, with strong national and system-wide support, a single entity that is legally empowered to review and approve data access requests on their behalf, according to shared common principles, with the detailed consideration that comes from a robust economy of scale in making a large number of decisions for a large number of settings, rather than a small number of reviews in a single organisation.

Addressing specific roles and uses

IG 21. Start an overdue public and professional discussion on performance management

The issue of data being used for performance management is informing a number of strategic choices around data access and analytics without open public discussion. Data plays a vital role in contributing to the improvement of quality, safety, and efficiency in public services. There should be a more robust and professionalised commons of knowledge around this work, as per the chapter on Modernising Service Analytics. Overall, this is a complex and also political area: it should therefore be the subject of a single piece of policy work by the NHS Transformation Directorate in consultation with the wider community to facilitate frank discussion. In advance of detailed work, the following outline principles are proposed, but only as subject to more detailed evaluation:

1. Legitimate skilled users from national government and national NHS organisations should be entitled to access NHS data for performance monitoring.
2. Organisations whose data is used in such projects should be carefully consulted.
3. Any concerns they raise should be carefully considered
4. Any concerns they raise ahead of analysis should be clearly recorded.
5. Any national organisation imposing costs or inconvenience on health services with metrics claimed to be substantially flawed or uninformative should be subject to expert review and possible censure by an appointed body with appropriate technical skills (such as the UK Statistics Authority).
6. Any prior track record of poor quality or misleading analytics should be considered when considering future data access requests; serious breaches should result in revocation of access to data from patients' records.

IG 22. Ensure DHSC can access data when appropriate

The management and delivery of NHS care is complex, changeable, and spread across a wide range of organisations. However, we have not encountered any convincing argument why analysts in the Department of Health and Social Care should not be entitled to a legal right to access and process health data for the purposes of operational research and policy analysis, as set out in the NHS Act 2006. A combination of technical and organisational factors means that DHSC analysts are often blocked from accessing data, or at least substantially delayed, which can have significant ramifications for timely policy development, ministerial advice, and operational improvement. Rules regarding the mechanism of access, for example only within a licensed TRE and only by accredited researchers, should be the same as they are for the rest of the system. If the broader access afforded by secure national TREs cannot resolve this problem then a review of legislation and regulation should explore other means by which DHSC can gain access to run analyses on patient data, without viewing individual patients' records directly.

IG 23. Start an overdue public discussion about commercial access

TREs make a profound contribution to commercial work with NHS data, because they completely separate two distinct issues: the protection of patients' privacy; and the wider ethical and political judgement about the appropriateness of commercial access. However, while addressing the issue of privacy - and detaching it from the wider ethical, political and strategic issues - TREs do not address those other issues. There is a need for a frank public discussion about commercial use of NHS data (with due consideration to the upcoming National Data Guardian work on Public Benefit). This should not be regarded as deciding all aspects, but should inform the decision. It should be executed through citizens' juries alongside other forms of public consultation, and include a frank and informative explanation of the key role that commercial actors play in innovation of medicines, services, and digital technologies; due recognition of their potential conflict of interest (as is already well recognised); and due consideration of the best means to mitigate that risk. This discussion should take place after the system has adopted TREs, so that the ethical or public preference aspects of data use are not confused with privacy challenges.

There is a very extensive array of detailed policy work with large teams in diverse settings across government around the best mechanisms for appropriate cost recovery, and commercial engagement; these arrangements sit outside the terms of reference of this review. Nonetheless it is the overall view of the chair (BG) that commercial use of data should be welcomed within a sensible legal and regulatory framework; that the risks of misuse, poor quality analytics, and COI are substantially present in non-commercial users of data; that all uses should be in a TRE that shares complete activity logs; that the NHS should negotiate intellectual property rights in any innovations derived from access to NHS patient records data.

IG 24. Negotiate co-ownership of claimed commercial innovations from NHS data

When a company develops an "algorithm" such as a risk prediction tool they typically apply existing tools, techniques and code libraries to huge, richly detailed health datasets that were collected at

great cost. The code libraries used for this work, such as those to implement random forest "AI" models, are themselves often open source. This is not to diminish the effort, application, creativity and problem solving that such work entails. However, it is a fact that the successful delivery of such an algorithm is driven in very large part by the availability of extremely detailed health data to drive the models. This data did not just appear in the lap of the NHS. It has been collected at great expense, over many decades: it has been created, curated, matched, enhanced by the collective effort of hundreds of thousands of clinical and administrative staff in the NHS, and tens of thousands of technical staff. It would be very wrong for the value in such work to be captured exclusively by the group that executed the code. The relative contributions of the underlying data, and the individual teams using it, will vary from project to project. We suggest that if commercial users approach the NHS or government seeking access to data to develop proprietary code or tools in this way then a negotiation should be conducted at the outset around the profit share between the NHS and the users; that these should be made public not solely for transparency but more to help the NHS get informative public discussion and feedback from technical and IP experts on whether they are managing the share correctly; and that all data management and curation code created during the project should be shared as open, to maximise network benefits, avoid the current repeated duplication of curation effort, and ensure that benefits accrue to the NHS even from projects that do not deliver a commercial output or attendant revenue. This data management code can be captured simply as all activity will be within the TREs. For clarity, this is a very different scenario to publicly funded code: this should be managed as above.

IG 25. Address exclusive commercial arrangements

This is an evolving space that has been the subject of various national policy initiatives over time, and substantial ongoing work across government. Overall, exclusive commercial arrangements are likely to be disadvantageous to the NHS as a whole. They will often represent local challenges at well-intentioned organisations: NHS staff have described feeling disadvantaged when negotiating with large commercial organisations, or even academic institutions, who may both have greater legal and commercial expertise on data access issues. The work of the NHS Transformation Directorate business unit 'The Centre for Improving Data Collaboration' should help ease this information and power-imbalance, by providing both 'off-the-shelf guidance for data partnerships, and bespoke advice. This work is important and should continue to be supported, underpinned by an acceptance that there is unlikely to ever be a 'one-size-fits-all' contract and so there will need to be a degree of pragmatism and contextual flexibility. Given the sensitivity of the topic, any guidance regarding commercial data partnerships should be informed by careful deliberate engagement with patients and different publics.

Patient and Public Involvement and Engagement

Good PPIE is crucial, and valuable. The team do not claim deep, detailed, technical knowledge around best practice in PPIE specifically. However, from the perspective of extensive prior engagement work - and multiple detailed discussions across the community of those conducting, using, and supporting research - we respectfully suggest that the following changes to the ways in which PPIE is funded, commissioned, and conducted could be considered.

IG 26. Ensure PPIE expectations are proportionate to the sensitivity and scale of the project.

Expectations from funders and regulators around PPIE can sometimes be the same for both large and small projects: this can be unrealistic, and may sometimes act as a barrier to younger and less resourced individuals or teams accessing funding and data. This is especially the case when funders require long and detailed PPIE work to be done in advance of any funding being acquired, at the application process. At minimum this additional cost prior to funding requests should be recognised and - if extensive pre-application work is deemed appropriate - this should receive specific resource that is accessible to all applicants at all levels, through all university departments; there should be due audit or exceptions reporting to determine whether this PPIE facility is available to all. More broadly: expectations regarding the quantity - not quality - of PPIE can and should be contextually flexible to the size of the project, and this should be clearly stated.

IG 27. Provide researchers with easy access to practical guidance, and examples of best-practice.

Good engagement requires deep and specific professional skills that may fall outside the skillset of data scientists or research software engineers. Ideally this skills-gap should be filled by embedding a social scientist, data ethicist, and/or engagement professional in the research team. This is not, however, always possible. PhD students, for example, do not necessarily have this option available to them. In these instances, it would be helpful for these researchers to have access to practical tools, resources, and guidance that can help them conduct high-quality PPIE. Examples of practical guidance do exist, but these can be hard to find for those who do not know where to look. Developing a central repository where resources can be shared and signposted to would help significantly.

IG 28. Resource and give a platform to experts in building public understanding.

Engagement as a two-way process is extremely important; but understanding is important too. Patients are entitled to a clear, adequately detailed, accessible description of what their data is actually being used for. The Understanding Patient Data website is an extremely strong example of good practice in this regard, with their diverse and thorough case studies at varying levels of technical detail. It is disappointing to see that this programme has recently been closed by Wellcome; it is to be hoped that the work will find a strong new setting.

IG 29. Consider centrally commissioning PPIE on common causes of concern

Multiple standalone PPIE projects on multiple individual analytic projects have a clear role, and strong support. However, for commonly raised concerns, and large topics, it may be useful to consider central commissioning of very thorough and detailed PPIE projects using agreed methodologies on challenging recurring questions in this space. This will help ensure everyone is able to benefit from the knowledge generated, and that the topics are given the level of attention, space, and consideration required.

Chapter 06

Data Curation

Goldacre Review

Summary

"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.

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. Furthermore, 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 location, is 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 pre-diabetes can also often be inferred from other traces on a patient's record, such as blood test results, treatments, referrals, or test requests. Lastly, 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, which will need to create single variables to denote concepts such as "patient has asthma" or "asthma review has taken place".

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”; and almost no formal culture of sharing, no 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.

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. This will help reduce duplication, build a commons of knowledge, and build capacity through reciprocal learning. Secondly, the system should create an Open Library where all NHS data curation work can be shared; with appropriate technical features as described below and in the full text; and an obligation for all curation work to be shared here. Thirdly, a small number of Data Pioneers should be resourced to populate this library with curation code on key clinical topics and areas. Fourthly, there must 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. Lastly, all curation work should ideally be conducted in standard TRE settings as this will inherently be more portably and re-usable code.

This 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 missteps 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.

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: 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.

Background

How raw NHS data is converted into a usable dataset

Before describing how curation can best be achieved, this section contains a brief overview of the practical technical reality, showing how NHS raw data is presented, and how it is extracted and transformed into analysis-ready datasets. This section can be skipped by those who feel they already understand this work, or do not wish to understand it; however, all technical and strategic initiatives to improve curation must start from a clear understanding of the underlying processes around EHR data curation.

As a simple example, it is useful to consider an analyst working with real GP records to create a variable for each individual patient denoting whether they have, or do not have, hypertension (high blood pressure). They could be an NHS service analyst, producing a report for local clinicians and commissioners that describes whether patients with hypertension are receiving the correct regular medication reviews, across all organisations in their local area, to the same extent; or they may be an academic researcher, aiming to produce an academic paper describing whether patients with hypertension are more or less likely to develop a particular kind of stroke.

As discussed previously, each patient’s GP record can be thought of as a series of events, each with a patient identifier, a date, a location, and an associated event code from a dictionary such as SNOMED-CT. Those codes could denote a diagnosis, a referral, a prescription, a test request, a test result, and so on. There may also be another variable associated with the event, such as the value of a test.

Patient ID	Event code	Associated variable	Event definition	Date, Time	Location
979384758	38341003		“Diagnosis of hypertension”	30/6/2021 10:31am	City Surgery, Birmingham B1 1AA
979384758	271649006	155	Blood Pressure systolic reading	30/6/2021 10:31am	City Surgery, Birmingham B1 1AA
979384758	VMP 318855006	28 tablets	Prescription for Enalapril	30/6/2021 10:31am	City Surgery, Birmingham B1 1AA

SNOMED-CT Code	Name
193003	Benign hypertensive renal disease (disorder)
1201005	Benign essential hypertension (disorder)
1474004	Hypertensive heart AND renal disease complicating AND/OR reason for care during childbirth (disorder)
6962006	Hypertensive retinopathy (disorder)
16147005	Arteriolar nephritis (disorder)
28119000	Renal hypertension (disorder)

The NHS as a whole across various organisations manages huge volumes of records such as these: approximately 50 billion rows in total (but with extensive duplication of records as discussed in the chapter on TREs). The analyst, however, does not need or want to know all this detail. For each patient in their analysis, they just want to know whether they do, or do not, have hypertension. This variable must be created by someone, somewhere, running analysis code across the raw records. For example, code may be written to check in each patient's record for the presence of any diagnostic code associated with hypertension. This is already no small task: there are hundreds of diagnostic codes that fit the bill for hypertension. Some examples are given in the table above.

From this list of codes the complexity of the job rapidly becomes apparent. There are many ways to record a diagnosis of hypertension in a patient's record. This is partly because there are many different varieties of hypertension (and partly because of other complexities and necessities around creating a versatile global approach to clinical coding). There are over 300,000 SNOMED-CT concept codes in total, across all concepts in medicine (these can then be combined into new codes, just as the German language can combine single words to make new and longer ones). Someone needs to search through this vast list and create the relevant subset of codes denoting that a patient has hypertension.

They might search in a SNOMED-CT code browser for any code matching the word "hypertension", and include all of those. But this job must be done attentively: it is not uncommon to see analyses where the list of codes for high blood pressure includes the SNOMED-CT identifier for "ocular hypertension", an eye condition that has nothing to do with blood pressure. Furthermore, not all hypertension codes will necessarily include the word hypertension. For example, "arteriolar nephritis" is in the list of hypertension codes above, with good reason: it is a form of progressive kidney damage caused by long-standing, poorly controlled high blood pressure.

Creating these lists is therefore complex and difficult work. To do it, the team need a number of skills which may not all be embodied in one individual. They will need domain knowledge about clinical medicine, to spot the need to include a code such as "arteriolar nephritis". They will also need domain knowledge about how health data works: they will need to know what SNOMED-CT codes are, how they can be searched, how codes are nested within a poly-hierarchy (and then challenging details such as how parent-child relationships in the poly-hierarchy can create overlapping conflicts, which can in turn change when individual codes move between nodes in different releases); they also need knowledge of how clinicians use clinical systems, what is recorded, how it is coded, and what the strengths and weaknesses of the data are (collectively this is often called "clinical informatics").

Then there will be some judgement calls. Some codes might be edge-cases for inclusion. This might depend on what the specific analysis is looking at, and so will entail the curator having knowledge of the full analysis plan, or the analyst having knowledge of how the variable was curated, so that each can spot problems (as seen in the Ethnicity codes example later in this chapter).

This complexity in turn means that it is often misguided to imagine that one might create the single, canonical, all-purpose codelist for "hypertension", as many have sought to create: some studies might want to include women with hypertension in pregnancy; some might not, for good reasons; some might want to be inclusive and avoid the risk of missing true hypertensives, at the risk of some false positives; some might want to use a more specific list; and so on. Canonical codelists have a role, but only for certain specific tasks, as discussed below.

This complexity also means that codelists are often imperfect, for good reasons: it is therefore very problematic that codelists are often not widely shared between individuals and organisations, either because analysts and academics don't think to share them, or sometimes because analysts specifically want to retain a competitive advantage over other teams, by having a good set of existing codelists to work with. Indeed, many publicly funded codelists, created by publicly funded organisations, are kept actively secret, withheld from the wider community, because the group that created them has the intention and the intellectual property rights to monetise them. The problems caused by these closed practices (and the wide system benefits of "buy out" models when the NHS pays for this kind of curation work to be done) have already been discussed in the chapter on Open Working Methods.

Then there is a deeper problem: many patients with hypertension don't have a diagnostic code recorded at all; but they do have multiple measurements of their blood pressure, where it is recorded as being systolic >140mmHg (which is high); or they are on lots of blood pressure medications. So now the person creating the variable for their analysis will need to come up with some rules, to create a "hypertension" variable, that rely on treatment codes, or measurement codes, with logic like "if this valuable is higher than 140mmHg". And they will need to do this for every diagnosis, or patient feature, that they are interested in.

Setting aside the complexity of making good codelists, there are then further problems. A codelist alone is rarely enough to create an assertion about a patient, such as "has hypertension". For example, should patients be described as hypertensive if they only have one diagnosis recorded, fifteen years ago, with no subsequent treatment? That might depend on whether they have been receiving any treatments subsequently; or it might depend on what kind of analysis is being conducted today; and so on. To work around this, the analyst doing the curation would need to write some very complex logic: a patient might be denoted as having hypertension "IF these diagnostic codes are present between 2015 and 2020; OR IF these diagnostic codes are present between 2000 and 2015 AND these treatment codes are present more than N times", and so on.

Then, where there are ambiguities, or edge cases, it might be useful to validate the variable created against another source. This could be indirect validation: for example, does the prevalence of hypertension for your variable match your expectations from other work (such as bespoke research surveys) on the prevalence of hypertension in this age-corrected population? Or it could be direct validation. You might even directly contact some clinicians and ask them direct questions about some patients you are unsure of, to see whether your approach to edge cases was sound. This latter kind of validation work is hugely valuable - and was done extensively with hundreds of staff in the 1990s when GP records were first being computerised - but has now fallen away to become a rare and niche activity.

"There is no one version of the truth because there is no true value of any characteristic defined by a measurement... What is the true number of the people in a conference hall? Do we count someone who temporarily leaves the hall to take a telephone call? Do we count the staff managing the audio-visual equipment? Do we count the people who join via the internet? Do we count the people in breakout rooms? If our job is to prepare lunch, then we only need to count the people who will stay for lunch!"

- Interviewee

The broader context

The examples above are drawn from work with GP data, which is widely regarded as the "jewel in the crown" of NHS data, but the principles are all directly applicable to all health and social care data. Hospital data is a valuable illustration: there are 2 broad categories of such data. At one extreme is SUS and HES, the extremely sparse datasets created by administrators to describe, in the broadest terms, the main purpose and date-range of each hospital admission. This sparse data is typically curated using lists of ICD10 codes, using simpler versions of the methods described above (although this is complicated by the fact that users of HES or SUS often receive very different versions of that data from each other, or different versions from year to year, because it is commonly somewhat pre-prepared within NHS Digital prior to release, adding additional layers of complexity and non-reproducibility between settings and between calendar years). At the other extreme, there is the full raw data residing inside hospital EHR systems, which has been largely untapped: this typically includes phenomenal degrees of detail, such as oxygen measurements taken many times a second, and more. Similarly, individual bespoke research data collections contain a huge variety of different data structures, coding schemata, and dictionaries - often bespoke - to capture the outputs from (in some cases) decades of questionnaires, tests and other bespoke data collections. The curation challenges for these various forms of data are phenomenal, and can only be addressed by taking a systematic and open approach.

How data management is currently implemented

During the course of the Review the team discussed the practical aspects of data curation work extensively with NHS service analysts, academic researchers, and private providers of health analytics. It is clear that this work is typically done in siloes, with minimal formal open sharing, to the frustration of many analysts. However, it is also clear that there is a vast amount of deep expertise currently left untapped and mostly undocumented, in individuals and teams across the system with domain knowledge around clinical relevance, the structure and meaning of individual data elements, the strengths and weaknesses of different national and local datasets, their provenance, and so on. This is captured in local documents at best, but often resides in the tacit knowledge of individuals and teams, and shared through informal personal relationships, because of an absence of any expectation or technical facility for wider sharing.

"Data curation is always on the back-burner."

- Interviewee

It is also fair to say that in many teams, and perhaps most, the practical implementation of data management is extremely ad hoc, with very different methods used in different settings. This reflects historic norms and is no criticism of individuals: the outstanding work done around Reproducible Analytic Pathways in other parts of government via GDS and ONS is comparatively recent. In NHS data science across both NHS service analytics and academic research there is very little evidence of a strategic approach to implementing these approaches, and very few teams turning local working practices into re-usable code.

Data management in the NHS today is therefore typically implemented in rather manual terms, as described in the section on RAP in the Data Curation chapter ((as the approach that RAP can replace). A reasonable thumbnail sketch, abstracted from numerous discussions, would be as follows: "Alice describes a clinical and commissioning question to Aaron; he describes the dataset he would like to use to Anna; she goes to their database, and writes a SQL query (or uses a point and click interface) to define the dataset to be extracted; this is executed, and the dataset is sent over to Aaron on the "shared drive"; Ada in his team then processes this a little more on this in Python, or R, because the dates can be done better in those packages; this is then sent to Aaliyah who does some matching onto another dataset in Excel; this is then saved on the shared drive, and turned into a dashboard using Tableau by Aaron. Some screenshots from this are included in a Word document report that is converted to a PDF and sent out to senior stakeholders for review. When the numbers need to be updated, the entire process must start from the beginning". In many cases, these different steps may be split between different organisations, introducing additional complexities, delays, and dependencies.

This portrait does not represent the very best of analytics in the NHS, and it does not represent the worst: it describes a common reality that is of its time, but would benefit hugely from strategic modernisation. Many individuals have very strong data science skills and domain knowledge; existing independent projects like the NHS-R Community have done excellent work to up-skill analysts for certain types of analytic work; and there are many isolated examples of individuals or teams taking an approach closer to that of Reproducible Analytic Pipelines (below). However overall, it is clear that the great potential from outstanding and skilled analysts across the system is not being harnessed due to a lack of coherent approaches, frameworks, skills, and tools in which they can operate.

Variation in implementation of data science

In the previous chapter there was an extensive description of RAP and modern, open, computational approaches to data management and analysis. While RAP is powerful, it does not fully address an additional opportunity to rationalise health data science and make it more efficient: the very substantial variation in approach to technical implementations for data centres in different organisations. More specifically: similar or even identical data is repeatedly held in different forms and different environments across the system in hundreds of different locations, including CCGs, STPs, Federations, PCNs, LHCRs, ICSs, CSUs, AHSNs, academic individuals, groups and organisations, private providers of health analytics, and more. The differences in implementation may be small simple changes, such as the types of database and SQL required to interrogate them, or different column headers, limiting simple portability of code; but these differences often expand into substantially divergent data models and over-arching approaches to storage and retrieval of information.

The problems caused by this diversity and lack of a coherent approach are manifold: it means that code even for isolated elements of a wider data management pipeline, such as a SQL query, is often barely useful to staff in another setting, and indeed barely intelligible, as it written bespoke to local bespoke database schemata and implementations. In some settings, such as those where higher volumes of analyses are implemented, there may be some more standardised or efficient approaches to data preparation: however there were no strong examples of this work being shared, and some of it being actively withheld, apparently driven by contractual approaches that permitted or incentivised data curation approaches being treated as high value intellectual property, if not explicitly, then via contracting that focused solely on the delivery of a final dashboard or report with no attention paid to the intermediate data management work.

This variation typically occurs for no reason other than an accident of history and implementation. Much of it is reasonable and of its time, as is the absence of sharing, as each implementation was inward facing and serving only internal needs, responding to system incentives. Unfortunately, this means the variation has been largely unnecessary and uninformative, because it has gone largely undocumented: there was little evidence of any openly accessible work describing different implementations, strengths or weaknesses, or even clear documentation for most settings, including for large and recent high value projects aiming to aggregate large volumes of data in local data centres for direct care and population analytics. This variation and duplication provides substantial opportunities for efficiency gains, and quality improvements, from national strategic approaches to modernisation in both working practices and the underlying technical implementation. It is compounded by one final historic overhang in the mechanisms commonly used for preparation of health data, which similarly presents a substantial opportunity for rapid improvement.

Requesting Datasets in Code

As above, raw datasets are transformed into analysis-ready derivatives, less disclosive derivatives, or required subsets of the full data, by executing code against databases containing the raw data. However, current approaches to data preparation prior to dissemination between organisations commonly entail a less transparent and efficient intermediate layer. Numerous interviewees told us, for example, that every time they wanted new data from NHS Digital based on SUS/HES they were required to have a series of interactions, in person or over email, in which the specifications of the dataset they wanted - a derivative of the underlying raw records held in NHS Digital - were discussed in conversations, or in free text. Sometimes these discussions were supplemented with lists of codes, but sometimes they were forced to rely on broader discursive conversations around the clinical or demographic characteristics of interest for the planned analyses, to be interpreted in-house by NHS Digital. Some users were comfortable with this approach as it required less expertise from their own analysts. Many expressed surprise and concern that they could not specify the dataset they wanted in code.

"Please, get them to tell me the database schema, tell me the codes and counts, I would much rather just write the SQL query for them to execute!?"

- Interviewee

This process of requesting datasets in conversation was described as creating a range of problems for users. For example, it creates substantial unnecessary work if datasets are not provided consistently when updated: one leader of an analyst team told us "every time we get a new cut of HES for our patients, even small things change, for example, the names of the column headers are different, meaning that all our data ingestion and matching code breaks, which requires extensive bug-hunting and rewrites for something that should just work."

More concerningly, this closed and conversational approach to dataset requests means that users cannot be certain that the data delivered meets their requirements. It prevents them being involved in understanding or contributing to the data management which is a core technical and informed element of the data analysis itself. It also risks errors that are all the more likely, and more impactful, because they are hard to detect. The team was told of cases where external validation through counts showed analysts

that the number of cases for a given condition identified by the methods used by NHS Digital or a similar dataset provider was substantially different from the known prevalence of the same condition in the same population, and therefore must be incorrect: this is not a criticism of any individuals or teams, but rather illustrates, again, that data management is a core part of delivering data analysis, and requires both technical general data management skills and domain expertise, which are both likely to lie to greater or lesser extents in both the database management team (in this case NHS Digital) and the analyst team addressing a given clinical or operational question. For high quality analytics, both should be able to develop, view and execute code across the while journey from raw data to final analysis.

"[I am] worried about what goes on in the black box, how did you do the linkage. There's [often] been egregious errors before the data gets to me."

- Interviewee

"We always say don't clean the data when it's done in Excel by just manually removing fields or changing entries -there is no traceability."

- Interviewee

A new systematic approach to curation

There are 2 core mechanisms by which the system can make data preparation work interoperable between settings. One is to work towards a harmonised data model and computational environment; the other is to accept substantial variation in implementation and environments, and work towards shared "portable representations" of data management code that can move between different data environments.

Each of these approaches is required in parallel: a harmonised model is realistic and desirable for some forms of data, especially national datasets such as GP data, HES and SUS, where variation in implementation is generally needless and unhelpful; while warranted variation will also be an ongoing practical reality, especially for very diverse local datasets such as local social care datasets, local service datasets, and bespoke local extracts from hospital systems. Tools and methods for both approaches should therefore be energetically developed, in the open, in a phase of rapid R&D. In both cases, this work will only be successful if it is combined with the following: an embrace of modern RAP methods; an expectation or obligation that all publicly funded data management code must be shared openly for re-use; and the provision of a curated library in which to share and discover such code alongside its documentation.

"It's really important to share. Even at a basic level. The number of times we've written HES queries that don't match up with someone else's HES query, and you're trying to align a cohort."

- Interviewee

Specific recommendations regarding how to develop a more systematic and efficient approach to curation are given at the end of this chapter.

The clinical importance of delivering data curation in code

At this point it is useful to return to the clinical relevance of this work. In particular it is useful to consider the enormous variety of activities where it is operationally and clinically important for different parts of the NHS and health ecosystem to communicate with each other clearly about patients and patient characteristics, but where this work is currently obstructed by the absence of a culture of communicating clearly in code, with portable representations as a common language. Good, clear, open, shared data curation is not abstract work: it is at the heart of all practical uses of data, and can directly impact on patient care.

In the detailed example of data curation above, there was discussion of whether a patient should be considered to have hypertension if they only have one diagnostic code in their record twenty years ago. This exact question is a live issue for clinical work that uses patient data directly in patient care. The CHA2DS2-VASc is a commonly used scoring system used to evaluate a patient's risk of having a stroke. Points in the scoring system are calculated based on the presence, absence and severity of each risk factor. Some implementations of this scoring system in popups in EHR systems used by doctors consider a patient to be a "yes" for "hypertension" if they have had two high blood pressure entries at any time ever in their record. Clinicians and analysts may reasonably debate whether that is appropriate. The issue is not whether any given single choice is correct or not: rather, it is whether there is a common language ensuring consistency so that when NICE or NHS England recommend to clinicians that they use CHA2DS2-VASc, or ask vendors to implement CHA2DS2-VASc in their EHR systems as popups, or as analyses of population treatment trends, all systems and analyses are using the same interpretation of CHA2DS2-VASc.

The team was told by senior leaders in informatics that it is a source of substantial concern to them that when they instruct vendors to implement a given decision support tool, or even a simple rule to generate safety alerts or popups for certain types of patient or activity, that there is ambiguity in the way that this is currently communicated. Similarly, vendors of EHR systems and other services that operate within EHR systems explained that verbal or discursive communication of complex patient characteristics was labour intensive, and often ambiguous or risky. There is a clear widespread need for open standards to communicate patient characteristics and clinical concepts in open standard code that is portable and can be implemented in multiple settings.

Similarly, when evaluating variation in health service activity or outcomes, it is vital that the same methods are used to identify given types of patients, and given characteristics within those patients, wherever these are implemented. As this work becomes more complex, standards and portable code become ever more important: NICE has expressed an admirable desire to work towards computable recommendations in their clinical guidelines, where appropriate, so that uptake can be monitored as a matter of routine; in many cases this will entail more complex clinical pathways and characteristics to be captured in code, and communicated between organisations. At present, NICE shares a range of recommended measures of patient care where the numerator and denominator are described in narrative text alone: the expectation appears to be that each local analyst in each local setting will interpret these narrative descriptions in their own way, in order to develop dashboards or tables describing local performance. Again, for simple and complex work of this kind, a common language - in the form of open standards for portable representations of data management work - is the only rational and efficient path to delivery.

It is impractical for different parts of the health and care landscape to communicate complex data definitions between each other in meetings and narrative descriptions rather than in code. All actors across the system should communicate about patient characteristics in "code not conversation". From discussions it is clear that some in the system may be reluctant to embrace this way of working because of concerns around accountability: in short, specifying patient characteristics in conversation leaves ambiguity, where it is someone else's job to implement it in code, whereas code that will be implemented in the real world feels definitive, final, and therefore worrying. This can be managed by resourcing the work adequately, giving adequate training, and developing mechanisms for quality assurance for code that will be implemented in critical clinical settings. Reducing ambiguity and developing accountability should be regarded as "a desired feature, not a bug". Concrete actions that can be taken to deliver portable representations of data management code are set out at the end of the chapter.

Understanding complexity in variables

There is a range of prior work on NHS data addressing tasks that are adjacent to the core challenges set out here around data curation, or that address a subset of the overarching curation challenge. This prior work includes some contributions on raw data cataloguing, codelist repositories, and curation of "approved" sets of indicators, codes or variables that are intended to be regarded as canonical. To understand the positive value of this work, and the gaps nonetheless remaining, it is helpful to first review some of the common challenges and complexities when curating EHR and other forms of NHS data. This is best done with a short series of examples.

Ethnicity

It is often crucial that analysts are able to know a patient's ethnicity for a wide range of challenges, including analyses that aim to evaluate inequalities in service access, or clinical outcomes, between different groups; and for studies describing the contribution of ethnicity as a risk factor for different diseases or clinical conditions. Ethnicity may be crudely conceived of as a single enduring feature of an individual, and for convenience, an analyst will generally want a single "ethnicity" variable or code for each patient in their analysis. It is common to hear analysts talk about accessing or linking "the ethnicity file". But the reality, and the codes in a record, are both more complex than this. Each patient might have their ethnicity recorded once, many times, or never, because ethnicity is considered in a patient's EHR record as an "event" like any other: a code, recorded at a given location, on a given date, during a consultation, or when first registering with a service (or when the ethnicity codes and categories used in the system change, triggering a need to update the coding for some or all patients).

Sometimes these codes will all match perfectly. Sometimes they will conflict, to a greater or lesser extent. For example, someone might be coded as South Asian in 2004, and Bangladeshi in 2014: it is reasonable to say that they are both, and that Bangladeshi is a subset of South Asian. In this situation the analyst may find that some patients have fine-grained ethnicity data, but not all. This reflects a range of common challenges in data management: is it always possible to map a smaller category to a larger one? Do the smaller categories move between larger categories over time? Then, to add to the challenge, individuals might move between groups in other ways: perhaps someone is categorised (by themselves in a form, or by someone else) as South Asian at one timepoint, and British Asian at another. This is a complex social and cultural question.

Which code should the analyst pick? The most recent? What if that conflicts with all the previous ethnicity codes? And what if there are starker conflicts, so substantial that they look like they might be errors, in one recorded ethnicity code? Clearly the analyst will need some rules to work through this: the logic to implement those rules will need to be captured in code, and that code will be far more complex than a simple codelist. Sometimes people call this kind of curation work "algorithms": this is a somewhat useful phrase, that should not be confused with algorithms used to calculate more complex concepts.

Then there is a broader challenge: different rules to curate and manage the data might be useful in one type of analysis, but harmful in another. Missing data is a good example of this. Approximately two thirds of all patients in England have their ethnicity recorded at least once in their primary care EHR data. That might sound low, but it is at least a representative sample, and that this data is missing at random: or rather, the proportions of ethnicity recorded match the proportions in the true population of the country, as ascertained by (for example) ONS census data. However, it is possible to improve the completeness of ethnicity data, because ethnicity is also commonly recorded in the SUS/HES hospital data. Matching this data into a dataset will generally achieve completeness of ethnicity recording closer to 95%.

In some cases, adding in ethnicity codes from hospital records will be very useful, as it reduces the number of people with missing ethnicity codes. But in some analyses this can be actively harmful, since the ethnicity data is no longer "missing at random", as it was in the GP records. Because the more complete records are created by using HES, the analyst now has more complete coding of ethnicity for the subset of the population that is sickest, and most likely to go to hospital. This can play havoc with some analyses, but for others, it makes no difference.

Childhood Asthma

In the example of hypertension, it was clear that creating a codelist for a diagnostic category can be a challenge in itself. But beyond this, there are additional problems. Let us take the example of a team trying to improve the care of children with asthma using data. They may be trying to produce triggers for "popups" in primary care EHR systems, or a dashboard to feedback to clinicians, or a national report, or a complex analysis of different care pathways and outcomes. To do any of this, they will need to identify children with asthma in the EHR records. Many children will have one of the many diagnostic codes for "childhood asthma" in their record. But there will also be many children who have regular prescriptions for specific inhalers, strongly suggestive that they have asthma. Should the analyst simply include all those patients? What if they just have one such prescription, suggesting a trial of treatment, to help unpick the diagnosis? What if they have a few, but very intermittently? What if the data curation code accidentally includes some patients who have a different condition, where the treatments overlap with the condition that the team are trying to identify in patients' records?

This complexity speaks to another recurring challenge when working with NHS data: analysts may want a range of different variables to describe the same clinical concept, that are appropriate for different work. For example, sometimes it is useful to have a very sensitive curation approach, that avoids missing true cases; sometimes it is useful to have a very specific curation approach, that avoids false positive cases. Again, this shows the need for a flexible approach to curation, where users can assert their own approaches, and where there is a rich ecosystem of informed judgement, kicking the tires, iteratively improving the work, and considering which approach best fits. This becomes even more complex when looking at complex categories such as "diabetes",

where there is no clear demarcation between diabetes and “pre-diabetes”, where coding and even treatment behaviour may vary between organisations and individuals, and where analyses that gloss over this complexity will risk generating incorrect answers and actions.

Again, this shows that there is often no canonical single variable for a given clinical or demographic concept, because different approaches might be more or less useful in different analyses. This is one reason why a library where people share, document, and validate variables is more useful than a canonical list. It also illustrates why this work should be done openly, and shared: it is laborious and difficult so it should not be done repeatedly for the exact same task; and it can be done well or badly, quickly or well, and if the methods are not visible to all, then shortcomings will likely go undetected.

Validation: Myocardial Infarction and Care Homes

When using this raw data, especially when the underlying clinical concept appears hard to capture, the sensible next step is commonly to “validate” the data. This can be simple validation at the level of a count of cases: “does the number of pregnancies that I see in my dataset, for the urban district I’m working on, match the number of pregnancies I would expect to see, given the age and sex distribution of the population?”. Or it can be individual validation, comparing different data sources’ characterisation of the same clinical concept for the same people.

Where validation is attempted systematically, it can often produce quite concerning results. For example, one study set out to identify which patients were recorded as having had an acute myocardial infarction - a heart attack - in each of 4 commonly used datasets: CPRD (a cut of general practice data for a subset of the population); HES (the hospital records set); MINAP (a national audit project collecting data on MI); and ONS death certificate data. The team had the same records, linked, for the same patients, from each of these 4 sources, in order to examine the overlap, or non-overlap, of MI records. Each data source missed between a quarter and a half of all MI events.

A 2021 study from my own group (BG) found similar problems when trying to identify, during COVID-19, which patients are currently resident in a care home. This is a common challenge for analysts across the system, and the data is needed by numerous teams to evaluate need and risk, and to drive preventive action. There are multiple potential data sources for this information, and each is imperfect. One might use household size and age profile, and infer that any address with a large number of people over the age of 65 is likely to be a care home. One might look at patients’ address data, and match this onto a list of care home addresses: there are multiple different ways to do this job, and some may be better than others; but they will all be unhelpful for the substantial number of patients whose address has not yet been updated in their records (especially for those having a shorter stay in a care or residential setting, for example after a hospital admission). One might use a SNOMED-CT code that GPs have recently been asked to use to denote whether each patient is in a care home, which guides some recent changes to GP payments: but it may go out of date; or it may not yet be complete for long term residents. Each of these methods produced very substantially different lists of care home residents, all strikingly discordant with each other.

“We end up spending a lot of our time correcting data that comes from other people. We are working on care homes - there are about 3 different sources, none of which actually tells you which ones are open and which ones are not. You have to do a lot of cleaning to get your actual insight. It’s mad to think about the number of people who are trying to do the same thing.”

- Interviewee

Studies such as these, on care home residents and MI cases, have huge implications for all analyses using NHS data: they speak to the crucial importance of a systematic approach to data curation, linkage, and validation. It would be a mistake to imagine that other nations’ data is better, or that other data sources are better: where these problems are not known, it is often due to a lack of looking.

Adjacent Contributions

At this point it is useful to consider work that is related to the challenge of data curation, but addresses only a subset of the problems. This is not to diminish the contributions, which in many cases are excellent and valuable: they are covered in brief to illustrate current art and progress; and to ensure there is clarity around the extant gaps.

Codelists

Codelists, as above, are an important element of data management, albeit that they do not represent - on their own - a complete description of how raw data is refined into an analysis-ready dataset (for example, they often require additional rules such as “more than one appearance of a code from this codelist within the date range month-year to month-year”, or more complex equivalents). There are numerous codelist repositories currently accessible online including: approximately 300 codelists created and shared by one research group in Cambridge; approximately 600 codelists from various academic teams at a codelist sharing repository run by an academic group in Manchester; approximately 700 codelists shared at an online repository run by one academic group in UCL; approximately 300 codelists in the SAIL “Concept Library” (which may mostly mirror the UCL content); and approximately 2125 codelists shared during COVID-19 from various research and commercial organisations at OpenCodeLists.org as part of OpenSAFELY. Some of these codelist repositories aim to address some of the larger challenges around implementation, annotation, or validation of codelists; but none do so completely.

It is useful to consider the broader context of codelists, as this area illustrates some of the challenges and opportunities around open ways of working, curation, and an emphasis on data science as a core legitimate activity separate from delivering single analytic outputs.

With certain strong exceptions there is only a rather limited academic research base on the process and hazards of developing codelists: this is an unfortunate oversight, and illustrates the limited investment and emphasis within the methods and research community on the foundational topic of data curation. Indeed, it is sobering to note that while policy discussions have considered enormous investments in Data Curation over recent years, conventional academic funders have barely assigned any resource to this work through their open competitive funding streams, and NHS service analysis contracts to commercial providers routinely allow all curation work to be done internally with no sharing or external site on the work.

There are also substantial challenges around sharing. Only a minority of the codelists produced are in an indexed repository; and many repositories are small, short-lived, limited, specific to an organisation, or otherwise limited, reflecting again the historic lack of emphasis and resource in this space. Sharing codelists also presents political and cultural challenges. Some analysts and teams are happy and keen to share their codelists. It is clear that others, by contrast, have been highly actively resistant: this has created some heated disputes within the community, including around whether work from groups reluctant to share was reliable, as it could not be reproduced or replicated without such information. As in the preceding chapter on Open Working, during our interviews we even encountered arguments that codelists could or should not be shared alongside academic papers reporting analyses because to do so would create legal liabilities.

In everyday practice there is also a passive failure to share, due to absence of incentives and structures to support such sharing. Sometimes codelists are shared as part of an academic paper, for example in an appendix, as a PDF document, but this can bring additional problems: these documents are hard to find, index, or webscrape for re-use, and the licensing or credit expectations are generally unclear. Often codelists are described as "available on request" – including in a notable recent output from a senior academic leader who speaks prominently on the need for open working. Sometimes they are only shared under a requirement for collaboration, which introduces further practical barriers to simple review, before even considering re-use. Commonly codelists are not shared, and sometimes they are regarded as commercially exploitable information: this is problematic, as there is little evidence of substantial commercial revenue, and it holds back efficiency gains and innovation through re-use and iterative improvement. All these challenges are driven in part because there is no clear agreement on rewarding, citing, or otherwise giving credit for intermediate features of analyses, such as codelists. These issues – and the need for better attribution models under a "team science" approach - are covered in the section on Open Working Methods. For the purposes of curation, in summary: codelists are shared; but only sometimes; without current structures or norms to facilitate and credit sharing; and this covers only a small part of the curation challenge.

Catalogues of raw data

"Data cataloguing" is describing the raw data and other core features, such as database schemata: for example, a catalogue might describe the long list of codes from a data dictionary, and their accompanying free text definitions, or a list of the possible variables that may be associated with it (for example, "age" will be defined as accepting a "floating point"; "number of children" as the type "integer"; and so on). This is important and useful work, that has been done excellently by organisations such as ALSPAC, one of many long-term "cohort studies" that have collected vast amounts of bespoke questionnaire and research data from thousands of volunteer participants: they share detailed catalogue information about their data, for their various users and potential users to explore and evaluate. Some but not all research groups share this information, where they are supportive of external use, especially where they have been resourced by funders to act as a common data asset. This cataloguing work is similarly already done for many of the very large datasets used commonly in healthcare, such as the primary care data dictionaries READ2, CTV3, and SNOMED-CT, or the ICD-10 codes used in SUS and HES, which are all well known to researchers working in the space. Within individual hospitals, there are myriad data structures and complexities to be explored (as below), and the cataloguing challenge goes substantially beyond a list of variables.

Creating this catalogue material is a substantial amount of work. There have also been various projects aiming to collect and index it in one place, including the HDRUK "Innovation Gateway". Several interviewees were critical of this project, saying that they wouldn't use it, don't think it's useful, and regard it as a modest low resource task. We disagree somewhat with these voices: this resource may be useful for a subset of potential analysts, such as those who are keen to get involved in conducting research, but have not yet developed the domain knowledge necessary to do so. However, the examples above make it very clear that a catalogue of raw data covers only one small aspect of the broader curation challenge; and un-prioritised or inefficient efforts to catalogue all possible data can risk displacing other higher value work.

Prior manual curation work

There is a range of prior work in the general space of data cataloguing, indexing, and codelists, which is useful and related to the broader goal of creating a coherent approach to data curation for analysis, albeit that it is often adjacent to the core task, or a subset of it. Within the limitations of scope, we attempted to conduct a brief overview of prior work in the UK. This was substantially hindered by historic norms around closed working practices, the locally specific nature of such work, and a tendency for it to be seen as lower value than the final delivery of a single analytic output.

Within the NHS, it is clear that there has been substantial work done under projects such as Local Health and Care Records, Integrated Care Systems, the National Commissioning Data Repository, and single one-off data aggregation projects such as Connected Cities North and others. This work has often been delivered by teams with deep knowledge around the meaning, interpretation, management and curation of data. It is likely that many of these projects - especially those with a deeper focus on NHS analytics - have delivered good work in private for smaller internal audiences, but it was not possible to find any accessible material. This provides a positive opportunity for discovery and re-use of the knowledge: any curation team created by the NHS should pursue this energetically.

Within academia from interviewees and desk research we became aware of various examples of substantial investment that was presented as focusing on data management and curation. However, in many cases these have not delivered a substantial body of open code, tools, documentation, or methods. Some senior and junior individuals in these teams, whose interest was more in data management and curation than in single academic analyses, expressed

strong concern to us about what they perceived as a diversion of resource to meet traditional academic metrics such as journal publication for single studies. Some were explicit that they felt resource, emphasis and effort in these data management projects had been re-directed away from the core stated purpose of the funding. They also expressed concern that this foundational work is not regarded by employers, funders, and journals as legitimate high value intellectual, methodological or practical contribution. These projects are not identified in this review in order to maintain a positive focus on future work. However, the challenge is compounded by a related problem expressed to us by various researchers with an international focus around foundational work in data analysis and curation: while funding for single academic research outputs is typically open and competitive, funding to develop methods, tools and outputs around intermediate tasks - such as methods and tools for data management or secure platforms - seemed to them to be closed, and accessible only in very large disbursements to a small number of incumbents. This is discussed in the section on TREs.

Methods and platforms for curation

Internationally there is a wealth of work that has been done, much of it open source, aiming to address various challenges around EHR data, federated analytics, and related challenges. This includes excellent work by Observational Health Data Sciences and informatics (OHDSI) and European Health Data Evidence Network (EHDEN), an international collaborative aiming to conduct epidemiological research across multiple countries by developing, implementing, documenting, and supporting open source tools, and training. There is also important adjacent work from highly complex programmes, both abstract and applied, around broader EHR challenges, including interoperability with projects like HL7, FIHR, OpenEHR, and the MBCK (Mobilising Computable Biomedical Knowledge) group, which engages with curation as part of its efforts to enable decision support tools to move between different computational environments.

Many of these projects provide outstanding examples of open collaborative working, good documentation, engagement with standards, community building, and the need to embrace the foundational aspects of data management and EHRs in order to deliver efficient and high-quality work.

Why libraries are better than standard variables

At this point it is useful to consider the commonly proposed - and tacitly used - approach of creating standard catalogues of "approved" variables and datasets, such as a single canonical variable for "patients with diabetes". This approach has several clear benefits. It can reduce some of the complexity in the system, and create useful shortcuts for analysts producing datasets. It can increase trust, if there is a sense that a given variable has been "assured" by a given organisation. It is also vitally important to create a common language for groups who all need to be sure that they are using common methods of ascertainment when talking about "patients with diabetes," or setting out to evaluate compliance with guidance or performance measures, where it is necessary to evaluate the proportion of patients with a certain condition who have a given intervention, or clinical outcome.

There is a substantial amount of good work that has been done in this space, much of it oriented around measures of performance in the NHS. Some of these are used in quality improvement projects, such as those run by organisations such as Health Quality Improvement Partnership (HQIP) and others, using national NHS data, local data, or their own bespoke data collections. Others are in Quality outcomes Framework (QoF), used to evaluate and financially reward GP activity. There are various prior efforts to gather, annotate, and discuss these variables, in numerator/denominator pairs, albeit in the form of PDFs that are hard to locate online.

The challenge comes when this approach is over-used, or used to the exclusion of a more flexible approach. If the system is excessively focused on developing and sharing singular methods to identify a clinical or operational concept in data, then this can block innovation and improvement in ascertainment and data curation, and obfuscate complexity. This is especially problematic in the common scenarios, as above, where the system needs more granular options within the over-arching concepts, that are best fitted to different kinds of analysis. It is also well known that variables created to measure performance can take on a political life of their own: they can create perverse incentives around coding behaviour for clinicians and organisations, and these in turn can compromise the validity of the data.

These variables, as components of performance measures, can also develop complex and powerful allegiance networks: by this they can become something closely guarded, protected, and resistant to change over time, even where they are substantial methodological shortcomings, or changes in context. They can also sometimes be controversial, and the topic of heated debate within and between organisations, at the stage of their creation, to the extent that they reflect certain organisational requirements or preferences as much as the best methods of ascertainment of a given concept in the datasets available. These pressures can mitigate against simple checks of validity, or improvement of the metrics. Where these variables are of high value to senior leaders in the system, or organisations with a stronger say in technical architecture, there is a further risk that data infrastructure and flows are built around the notion of a small number of standardised variables, which again tends towards inflexibility, and mitigates against dynamic innovation in clinical phenotyping, data management, and new analytic windows or questions.

In short, it is important that the system supports good quality work to create, spread, and "assure" individual variables that are used to meet specific operational needs, in specific analyses, for specific users in the system; but equally it is important that this does not become the only form of data curation, and that the shortcomings in this approach are managed effectively. This, combined with the plain complexity of data curation as outlined above, underpins the emphasis in this review on the need for a library: a place where variables can be stored, shared, indexed, curated, discovered, annotated, discussed, validated, technically documented, and improved upon. This is the only approach that can reflect the complexity in the system; the need for more systematic methodological innovation in characterising concepts in data through curation; and the complex diversity of end-users' needs.

Clinical Informatics and Raw Data Quality

Throughout the review interviewees described raw Electronic Health Record (EHR) data as "messy". This is understandable, but in some respects unfair. A patient's health record is, historically and by design, an aide memoire: a practical record, created by clinicians and patients, to help them manage the patient's care together. EHR data was not created for research or analysis, it was not funded as an analytic resource, and where there is a desire to use it for those new purposes, then it is the users' job to do the work to convert it into something that reflects their additional needs.

However, it is also clear that the creation of raw data, especially data coded directly at the point of care by clinicians, is a very neglected area of activity, research, and improvement. GPs collectively, in their daily work of seeing patients, create the vast national asset of the GP EHR dataset, but receive almost no formal training around how to collect data, and how to capture the issues, symptoms, diagnoses and treatments they see and use in structured data. We were told repeatedly throughout the review by clinicians that any training is typically ad hoc, on the job, specific to systems and organisations, and highly variable between settings. When coding systems or dictionaries change - from READ2 to CTV3,

and to SNOMED-CT - again, GPs typically receive minimal formal training.

This is a relatively recent phenomenon, and reflects the issue falling from sight in the policy and professional communities. In the 1990s, when GP practices were first being computerised (96% being fully computerised by 1996) there was a workforce of hundreds engaged to improve coding quality, employed by organisations such as PRIMIS. These teams would travel out to GP surgeries, explore their coding behaviour, examine the extent to which they were using different types of code, identify gaps and opportunities to collect better data, and support clinicians to improve data quality.

This workforce would best be characterised as working at the practical end of "Clinical Informatics": the applied science of how health information is captured, stored, represented, reshaped, re-purposed, and retrieved for use. This is a thriving field in many countries internationally, especially the US (where a large amount of detailed work with data is well resourced due to its value in managing complex reimbursement systems for private insurers). A very large body of interviewees, senior and junior, across a range of sectors, were extremely concerned that Clinical Informatics has been neglected in the UK in recent years, and that what persists of the field focuses less on deep technical skills and knowledge, and more on leadership issues, such as managing organisational change when new EHR systems are deployed.

It is certainly clear that clinical informatics is currently low status and low activity for training, funding, and research. Most clinicians have received almost no formal teaching on clinical informatics in undergraduate or postgraduate training. While some work may go under other names or brands, there is almost no evidence of open competitive grants available in clinical informatics from national funders, very few departments calling themselves Clinical Informatics in medical schools, and very little active research by comparison to other fields.

There is also a dearth of research in this field, with almost no work on variation in coding behaviour between clinicians, the best ways to improve coding quality and completeness, or validity checks of coded data; and numerous and large related areas of activity that seem foundational, but have seen almost no work, with substantial tracts of digital health activity poorly evaluated, understood, and described. "Popups" are windows triggered in EHR systems to remind clinicians of the need for a given action: these are created in vast numbers, with triggers and content set by individual clinicians or local staff, or purchased from vendors; but there is almost no work describing what popups are triggered where, comparing different approaches, iterating the best design, exploring "popup fatigue", how often they are dismissed or disabled; and their impact on patient care has been studied at best in expensive one-off projects to evaluate a single popup as if it were a single standalone intervention for an RCT, rather than one example from a class of many thousands in common use. Similarly, "templates" are commonly used to guide the collection of data in EHR systems, but there is almost no systematic work on their impact, what the best design of a coding t-emplate would be, how they can best be shared and re-used as standard data collection tools, and so on. Many researchers working in EHR data - especially those without direct and recent clinical experience - are unaware that such a thing as a template for structured coding even exists in EHR systems.

"It feels like when people talk about data science it's always the whizzy stuff, we need to do the basics [which is often harder], really basic stuff just doesn't work."

- Interviewee

Clinical informatics is therefore an area with huge potential, sitting at the nexus of multiple critical strategic activities for the NHS and the nation including healthcare quality improvement, data science, software development, open methods, data infrastructure. A range of practical steps to address this are proposed at the end of the chapter.

Data Curation as the first step to wider interoperability in the NHS

It is clear that data curation is crucial, foundational work. It is complex, but nonetheless rapidly amenable to the working practices that have been established and proven in other sectors to manage complex collaborative work on technical challenges by multiple teams: specifically, the use of modern open working methods, sharing re-usable code, harmonising approaches, documenting work, and maintaining good knowledge management in libraries of information.

Cautious strategic investment in this space is critical, and will achieve multiple objectives. Specifically, it can:

- Eradicate expensive duplication of effort
- Allow competitive innovation on higher value work
- Make individual analyses faster to deliver
- Improve the quality, accuracy and validity of analyses
- Act as a focus, spine, or coral reef for more detailed work on clinical informatics
- Help the life sciences sector understand NHS data and identify opportunities to innovate

It can also support a longstanding ambition in the NHS, re-iterated in the draft Data Strategy and elsewhere, to "separate the data layer from the application layer": more specifically, the plan to create a rich competitive and collaborative ecosystem where the underlying patient data is kept in a separate service layer; and where EHR systems, decision support tools, and all other services aiming to engage technically with NHS data and logistics can all access that data on an equal footing, without requiring any specific

single commercial relationship with specific local vendors (who are regarded by some, not always correctly, as asserting a monopoly over data access). This is an extremely ambitious project, that would entail the creation of (and compliance with) a truly monumental set of open standards, alongside extraordinary and unprecedented technical work to deliver a service layer capable of supporting responsive EHR systems across the entire country, delivering data to clinicians within milliseconds of its request.

By contrast, the speed of data flows required for analytics are substantially slower. In short, where there is any ambition to deliver this kind of interoperable future - which should be supported and admired - then the open code and social infrastructure for data curation is the ideal and necessary place to start. “Interoperability of code for analyses on EHR and other health data” is a subset of interoperability for all EHR and software systems in health and social care; while “systematic data curation” subsumes many of the same technical challenges around making diverse data models commensurable; and both require addressing many of the same social and cultural challenges around interoperability and standards. Curation’s merit, as the key place to begin this broader project of interoperability, is that it is tractable. It can be approached pragmatically and productively, in a part-wise fashion, addressing smaller aspects of the full challenge but nonetheless still delivering useful code, infrastructure, knowledge, working practices, and tangible outputs such as analyses and efficient analytic services.

This should be used as a positive challenge and proving ground: if the system cannot deliver the open code and social structures needed to deliver on data curation, then it is unlikely that the system can deliver substantial progress on the larger challenge of interoperability more broadly.

Recommendations

Summary

Data curation is a vast but crucial challenge, if the nation is to capitalise on NHS data to improve care. Data curation is a data management task. Data management is done in code: where there is a desire to share access to curated data, this means sharing access to re-usable data management code with adequate technical documentation, in a library where it is discoverable and managed. The system has previously considered very substantial investments in this space, reflecting the scale of the opportunity for analytics, research, and the life sciences. Every project with NHS data entails a huge amount of manual curation effort: every time this curation work is done without being captured, that is good work wasted. Progress can be made in a staged manner with the correct working approach, using the following principles and proposals:

- Adopt the principles and practices of RAP and computational methods for all data curation and analysis work, as per the Open Methods chapter.
- Develop and maintain pragmatic standards to ensure that data curation code is portable, working towards: the use of national TREs wherever possible; the use of standard local TREs rather than different data environments in every data centre; standard implementations of national datasets in local TREs; standard functions for data curation; portable representations of data curation actions.
- Generate data curation code:
  - Adopt the principle of “curate and share as you deliver analyses” to automatically prioritise curation work and ensure it is useful in real analyses.
  - Ensure all national and standard TREs facilitate code sharing and require all curation code to be shared.
  - Ask and, where appropriate, oblige data users to share curation code in portable re-usable forms, as soon as environments support this, into the Library.
- Create and maintain an NHS Data Curation Library where all data users can assert variables and data management code alongside technical documentation; this must be inclusive, and separate from work around assurance of variables.
- Create open competitive funding for key tasks:
  - Fund the development of standard data management functions in open code, and portable representations of data management, using open competitive funding calls open to all.
  - Applied methodological work and code on core activities around data validation at scale, sense-checking at scale, data description, portable representations, clinical informatics and curation.
  - Deep dives to deliver curation, validation, sense checking at scale, and data description for 1-3 single clinical topics, in projects co-led by practitioners and developers.
  - Surface prior curation work in the Library.
- Start with “Data Pioneers” sharing open code for curation, as part of their RAP work, in RAP compliant TREs, in ICS and national teams, to a prototype library.
- Oblige the system to communicate about patient characteristics in “code not conversation” when specifying datasets, patient alerts, and similar.
- Develop capability in clinical informatics through clinical undergraduate and postgraduate training, and standalone career pathways.

Cur 1. Adopt the principles of RAP

This is covered in the Open Working chapter, and is crucial. Data curation is a data management task. Data management is done in code: where there is a desire to share access to curated data, this means sharing access to re-usable data management code with adequate technical documentation, as per RAP. This must be more than an ethos: teams and individuals looking to change their way of working must be provided with training, tools, and platforms that support these working practices.

Cur 2. Set up an NHS Data Curation planning and delivery team, or similar, to own the challenge

Regardless of the scale of work planned, the system has longstanding ambitions and needs in this space. A single source of knowledge and oversight on curation can provide open documentation on current state of the art, run deep dives on single topics, oversee investments from across the system, and identify opportunities. This should coordinate initial deep dive and ideas meetings with those parts of the community with existing expertise including: AphA, NHSR Community, academic groups, local and national analyst teams, NCDS, MCBK, NICE, EHDEN, OHDSI, and more.

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

The NHS should commit to produce and maintain an open public library of all data curation code, accompanied by appropriate technical documentation, that can be populated by any user of NHS data. This library should have the ability to store curation code, associated with a range of additional information including documentation and information validity checks; and provide facilities for annotation, “forks” (a technical term for derivatives of code), and citation or other forms of credit and tracking

of use and derivative use. Code should be shared on a “user beware” basis, but capture the provenance of each entry including the user or organisation, so that those evaluating prior curation work can use this - alongside technical documentation and validity checks - in considering the extent to which they are happy to trust and re-implement existing code. It is crucial that this 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. This library should be maintained by a small team with domain knowledge around NHS data and its curation or use, attending closely to curation of collections, improvement of documentation, commissioning and evaluation of validity checks on higher value variables, and meeting library users’ needs. This team should engage in close informative 2-way dialogue with platform and curation teams.

Generate NHS data curation code, and surface existing work

Cur 4. Mandate that all publicly funded data curation code is shared openly

All data curation code written using standard tools against standard datasets in standard computational environments should be shared, mandatorily, for all publicly funded work, including all academic research, and all NHS service analytics, whether delivered by public or commercial organisations.

Cur 5. Identify five Data Pioneer teams to adopt open curation methods

These should come from a mix of local NHS analyst teams, national NHS analyst teams, and academic teams, with prior evidence of working to RAP principles (or strong potential to do so), who can work to RAP principles in a TRE that supports RAP.

Cur 6. Ensure national programmes lead by example

National data analysis programmes including QoF and the key national “variation in care” audits such as Model Health System, GIRFT and RightCare, all use national datasets, as well as some bespoke datasets, for their regular data reports. They are in a good position to lead by example on adopting new RAP working methods with standard data management tools in standard shared environments.

Cur 7. Capture, and openly share, existing curation knowledge around commonly used national datasets

National datasets such as SUS/HES and GP data are commonly used: there is extensive knowledge and best practice from local and national teams around converting raw data into usable variables including in NHS Digital, the work done on the National Commissioning Data Repository, many national audits, and some academic projects. This should be identified and captured in shared, re-usable curation code and portable representations alongside technical documentation and any existing validity checks for each variable. To ensure delivery this work should be done under the aegis of one team tasked with identifying and surfacing best practice from these communities in the NHS Curation Library.

Cur 8. Use consistent environments to facilitate re-usable curation code

Currently national datasets such as GP data and HES/SUS are stored and made available in many very different ways in a huge number of different national and local data centres. The system should aim to minimise variation wherever possible. This means minimising the number of locations where data is stored. Where multiple locations are needed, the same health data should always be stored in the same way; and it should be interrogable through a consistent computational mechanism.

Cur 9. Require use of national TREs for tasks using national datasets wherever possible

This is likely to address a large number of local analyses, especially those aimed at benchmarking local activity against national activity.

Cur 10. Create and enforce consistent standards for local implementations of national datasets

Wherever possible the same health data should always be held in the same form wherever it is stored or made accessible, not the many slightly different cuts, data models, or even column names that are seen across different settings. This will mostly require that all data, but especially national commonly used datasets such as GP data or SUS/HES, is held in in TREs, in its rawest forms, closest to what is collected at the coalface. Doing this will help to address the needless and uninformative heterogeneity of current data structures that obstruct code sharing.

Cur 11. Create and enforce standards for local TREs

Local TREs should be either one single open source NHS TRE design, or conform to an extensive range of open standards, as per the TRE chapter (recognising that developing such standards in isolation may be more complex than developing a template open source TREs for local use). Doing this will help to ensure that code and skills are portable between teams and settings.

Develop tools to facilitate re-usable curation code

There will always some duplication of implementation of the same datasets in multiple locations. Two modest actions can ensure that data curation code is portable settings.

Cur 12. Develop standard tools to convert raw data into analysis-ready datasets

The conversion of raw data into analysis-ready datasets should be done with common tools, regardless of setting, to ensure that all data management code is intelligible and re-usable by others. This will require the rapid development of standard functions and libraries (re-usable code and tools), most likely implemented in python or R, by a small range of national experts in collaboration with a broad group of technical users representing a wide variety of data curation and analysis needs, supporting command line and GUIs, that can be implemented in diverse settings. This will be radically more

straightforward if all local settings can be obliged to adopt a standard local TRE model: ICS's are a new set of organisations in the system, using data to improve care, and provide the perfect opportunity for such standardisation (see TRE chapter).

Cur 13. Develop portable representations of data management code

There is a need for portable representations of data management actions, as described above. This will ensure that curation code is readable, understandable, re-usable, and portable between teams and settings, and let organisations communicate dataset specifications, or clinical popup and alert specifications, between them. This work must go beyond mere codelist sharing, and address complex phenotypes. It should be led by technical teams, and coordinated by the NHS, specifically the NHS Transformation Directorate; delivery should be led by a team of individuals with proven prior expertise in technical aspects of data management, open standards, informatics, and - lastly - health data management. This should start with a minimum viable product addressing the simplest variable types; and any standards should permit collaborative extension into more complex variables. The latter will require open competitive funding through traditional research funders to develop methodological work and applied code. Work should begin by rapidly delivering a detailed overview of prior art in this space including the related open work done for interoperability around projects such as FIHR and HL7, the excellent example of productive open working practices embodied by EHDEN and OHDSI, and evaluate opportunities in the complex work done for OpenEHR.

Cur 14. Run an open competitive funding call for foundational work on data curation

Data curation is a complex methodological challenge. There are simple aspects of the work that can be met with simple "implementation" in code. There are also more complex aspects that require the development of new methods and abstractions. NIHR and/or UKRI should rapidly develop open research calls on a range of prioritised areas including work to address the following challenges:

- Describing the quality and completeness of coding on key clinical areas.
- Describing variation in coding behaviour between settings.
- Developing methods and code for validation and description of data at scale.
- Developing and evaluating interventions to improve the quality of coding, focused on specific clinical or geographical areas.
- Developing optimal methods, tools, and training for codelist creation and related curation tasks.
- Developing and implementing optimal methods for portable representations of complex clinical and demographic phenotypes.

Any funding should also invite other innovative approaches to developing better curation of NHS data, but require a clear pathway to implementation in real NHS data analytics within two years. In order to access the best expertise and ideas, and surface the largest amount of prior knowledge, it is vital that all funding is open to applications from all, rather than granted to a closed organisation or group; and that resource is available on realistic timescales (6 months delay from award to commencing work; minimum two years resource). This is crucial, as informatics is a historically neglected space with pockets of excellence and innovation that are under-resourced, and will need time to scale as with other areas of innovative research.

Cur 15. Insist that all dataset requests are made in code

Dataset requests should be made in code, not conversation. Data providers such as NHS Digital must be capable of receiving and supporting such requests; requesters must be capable of making them. The process of making a request from a data provider such as NHS Digital should entail writing in code the characteristics of the dataset requested for preparation, and how it is created from the underlying raw data. The public log of datasets released should include this code, not just a free text description of the project and dataset requested. This should be a non-exclusive arrangement, with the prior method of discursive dataset requests persisting in parallel, but such requests should be met by writing open standard code in-house which is then shared as with all other dataset request and preparation code. Delivering this will require that organisations providing data such as NHS Digital also provide well documented details of their underlying data models, datasets, data dictionaries, and so on.

Develop capability in Clinical Informatics

Cur 16. Ensure there is clinical informatics training on medical school, post-graduate, and other clinical curricula

Clinicians enter and use clinical data about their patients. There is a need for adequate core training on the purpose and importance of this work, how clinical data is stored, and how it is used for analytics and research. Identify and resource existing organisations such as Faculty of Clinical Informatics to develop training in undergraduate and postgraduate curricula, with openly accessible online training for those out of formal training.

Cur 17. Ensure universities have core capacity in clinical informatics

Evaluate the best means to develop core capacity in universities, capitalising on any training work to embed practical and theoretical informatics research alongside it.

Cur 18. Support core capacity in clinical informatics

There is a need for profession-building in this space. As an example, the Faculty of Clinical Informatics is small and funded by members. This limits its impact. It is unrealistic to expect individuals to resource the development of professional structures to meet national strategic needs: this should be supported by core investment.

Chapter 07

Strategy

Goldacre Review

The system as a whole has huge potential. NHS data is unparalleled in its breadth, 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.

The system has 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

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).

Build impatiently, but incrementally

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.

Identify a range of "data pioneer" groups from each key sector

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.

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.

1. Develop a common "service wrapper" for TRE access, with civil servants.
2. Develop common working practices for the "generic compute and database layer" of TREs with generic skilled technical teams from private and public sectors.
3. 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.
4. 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.
5. Ensure all TRE teams work in the open, sharing and documenting all code and working methods as they go, to support adaptive innovation.
6. 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).

Focus on platforms

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.

Chapter 08

Conclusions

Goldacre Review

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.

Professor Ben Goldacre, Declaration of Interests

Ben Goldacre leads a research group working in applied data science at the University of Oxford (bennett.ox.ac.uk). All outputs including all tools and analyses are open source and open code, free for review and re-use by all. Some of this work involves working with NHS England and NHS Digital services and / or data. Major projects include OpenPrescribing.net, OpenSAFELY.org (discussed in the review, with a further COI statement where this is discussed), OpenCodeLists.org and TrialsTracker.net. A full list of current papers and projects is at https://www.phc.ox.ac.uk/team/ben-goldacre BG's group has received research funding via the University of Oxford from a wide range of public and charitable funders: the NHS National Institute for Health Research (NIHR), the NIHR School of Primary Care Research, the NIHR Oxford Biomedical Research Centre, NIHR Applied Research Collaboration Oxford and Thames Valley, UKRI / MRC, NHS England, the Wellcome Trust, the Good Thinking Foundation, Health Data Research UK (HDRUK), the Health Foundation, the Laura and John Arnold Foundation, the Mohn-Westlake Foundation, and the World Health Organisation. BG also receives personal income from speaking and writing books (and, infrequently in recent years, journalism) for lay audiences on the use and misuse of science and evidence.

Acknowledgements

During the review we conducted more than 200 small-group or individual interviews as well as eight open focus groups, and received over 100 written submissions. We are grateful to everyone who contributed their time, thoughts, and ideas: without them this review would not have been possible. The individuals and groups listed below (in alphabetical order) have consented to being publicly acknowledged for contributing to the review in some capacity. Their involvement in this capacity does not imply that they are responsible for, or in agreement with, the full contents of this Review.

Individual discussions and small groups

Thank you to the following individuals for giving up their time to write to us and/or participate in an interview with us.

1. Dr. Abigail Emma Russell, University of Exeter
2. Dr. Achim Wolf, NHS Test and Trace, DHSC
3. Mr Adam Manhi, Flatiron Health UK
4. Mr Adrian Jonas, NHSE&I
5. Aidan Peppin, Ada Lovelace Institute
6. Dr Alan Mighell, Dean, School of Dentistry, University of Leeds
7. Alastair Cartwright, Informatics Director (Leeds Health and Care partnership), NHS Leeds CCG
8. Dr. Aleksandra Gentry-Maharaj, MRC CTU at UCL, UCL
9. Alex Bailey, Programme Manager, Medical Research Council Regulatory Support Centre
10. Prof. Alex Elliott, University of Glasgow (Emeritus Professor of Clinical Physics)
11. Dr. Alexander Renziehausen, National Cancer Research Institute (NCRI)
12. Professor of Perinatal Health, Alison Jill Macfarlane, University of London
13. Alison Pritchard, ONS
14. Dr. Amir Mehrkar, GP; University of Oxford
15. Miss Amy Davies, The University of Bristol
16. Andi Orlowski, Director, The Health Economics Unit
17. Mr Andrew Davies, ABHI
18. Mr Andrew Eland, Diagonal
19. Andrew Engeli, Deputy-Director for Innovation and Partnerships,Joint Biosecurity Centre
20. Prof. Andrew Farmer, NIHR Health Technology Assessment Programme / University of Oxford
21. Prof. Andrew M McIntosh, University of Edinburgh
22. Mr Andy Boyd, University of Bristol
23. Mr Andy Wheeler, Getting It Right First Time
24. Dr Angela Coulter, Public Advisory Board, HDR UK
25. Dr Ania Zylbersztejn, UCL Great Ormond Street Institute of Child Health
26. Anita McGrogan, Senior Lecturer, University of Bath
27. Anna King, Health Innovation Network & DigitalHealth.London
28. Dr. Anna Price, Office for Statistics Regulation
29. Ms Anne Marie Morris, MP, House of Commons
30. Anne Mason, Centre for Health Economics, University of York. Submission on behalf of ESHCRU II (the NIHR Policy Research Unit in the Economics of Health Systems and Interface with Social Care)
31. Dr. Anoop Dinesh Shah, University College London
32. Dr. Anya Skatova, University of Bristol
33. Prof. Ara Darzi, Imperial College London
34. Ayub Bhayat, Director of Insight and Data Platform, NHS England and NHS Improvement
35. Dr. Balint Stewart, DHSC
36. Barry Sandison, Australian Institute of Health and Welfare
37. Dr. Becca Wilson, University of Liverpool
38. Benjamin Kelly, Director of Research, Outcomes & Data Science, Nuffield Health
39. Mrs Bevleigh Evans, NHSEI
40. Brian MacKenna, NHS England & The Bennett Institute for Applied Data Science
41. Dr Bronagh Walsh, University of Southampton
42. Bruce Richard, Consultant Plastic Surgeon, Birmingham Women's and Children's Hospital, Cleft Research Charity
43. Prof.Calum Semple, OBE, University of Liverpool
44. Ms Camilla Ravazzolo, Market Research Society
45. Prof. Carol Dezateux, Academy of Medical Sciences and Queen Mary University of London

46. Caroline Cake, Chief Executive Officer, Health Data Research UK (HDR UK)
47. Catherine Ayland, Ethics Advisor, British Healthcare Business Intelligence Association (BHBIA)
48. Dr. Catherine Bromley, Economic and Social Research Council
49. Catherine Pollard, Director, Centre for Improving Data Collaboration, NHSX
50. Dr. Catherine Saunders, University of Cambridge
51. Dr. Charlie Davie, DATA-CAN: The Health Data Research Hub for Cancer
52. Ms Charlotte Johnston, AbbVie
53. Chris Mullin, Chief Analyst, Department of Health and Social Care
54. Chris Wigley, Chief Executive, Genomics England
55. Dr. Christopher Bunch, UK Caldicott Guardian Council
56. Prof. Christopher Holmes, The Alan Turing Institute, and University of Oxford
57. Ms Claire Palmer, King's College Hospital NHS Foundation Trust
58. Mr Conor Briant, Kent Surrey Sussex Academic Health Science Network
59. Cory Doctorow, Special Advisor, Electronic Frontier Foundation
60. Damini Satija, Senior Policy Advisor, Centre for Data Ethics & Innovation
61. Mr Dave Buckley, Centre for Data Ethics and Innovation
62. Mr David Boothroyd, SCW CSU
63. David Dodwell, Consultant in Clinical Oncology, National Audit of Breast Cancer in Older Patients
64. Prof. David Ford, Population Data Science, Swansea University Medical School
65. Prof. David Forman, University of Leeds
66. David Sibbald, CEO, Aridhia
67. Mr David Snelson, use MY data
68. Prof. David Spiegelhalter, University of Cambridge
69. Dr David Stables, Endeavour Health Charitable trust
70. Dr Douglas de Jager, Human.ai Limited
71. Dr. Brian Roberts, Head of the TRE Service for England, NHS Digital
72. Dr. Bryan R Deane, New Medicines & Data Policy Director, Association of the British Pharmaceutical Industry
73. Dr. Chris Bates, Director of Research & Analytics, TPP
74. Dr. Fred Kemp, Deputy Head of Licensing and Ventures, Oxford University Innovation
75. Dr. Gurpreet Singh, Chief Health Science Officer, FITFILE
76. Dr. John Parry, Clinical Director TPP
77. Dr Marc Farr, Chief Analytical Officer, East Kent Hospitals NHS FT
78. Dr. Nicola Stingelin-Giles, Ethics Researcher, Royal Medical Society, Medicine & Society Section
79. Dr. Richard Fagan, Director BioPharm, UCL Business Ltd
80. Dr. Nathan Hill and Mr. Jagtar Dhanda, Bristol Myers Squibb
81. Prof. E. Richard Gold, McGill University, Faculty of Law
82. Ed Humpherson, Director General for Regulation, Office for Statistics Regulation

83. Dr. Ed Smith, The Christie NHSFT
84. Elizabeth Gaffney, Head of Data Access, NHS Digital
85. Miss Elizabeth Waind, ADR UK (Administrative Data Research UK)
86. Ellie Eastwood, Health Innovation Network
87. Ellis Parry, Data Ethics Adviser, Information Commissioner's Office
88. Dr, Emma Gordon, ADR UK programme, ESRC
89. Miss Estelle Spence, NHS Digital
90. Prof. Ewan Birney, EMBL-EBI
91. Dr. Farah Jameel, British Medical Association
92. Fran Woodard, Executive Director Data and Analytics Services, NHS Digital
93. Dr. Francesca Cavallaro, UCL Institute of Child Health
94. Mr Frank Wood on behalf of Leeds Informatics Board
95. Dr. Gareth Price, The Christie NHS Foundation Trust
96. Mr Gary Leeming, Liverpool City Region Civic Data Cooperative
97. Dr, George Millington, Academic Vice-President, British Association of Dermatologists
98. Glen Robinson, National Technology Officer, Microsoft UK
99. Glyn Jones, Chief Digital Officer, Welsh Government
100. Gordon Adams, Strategic Intelligence Manager, Salford City Council
101. Guy Cohen, Privitar
102. Dr. Helen Firth, Wellcome Sanger Institute
103. Helen Louwrens, Director of Intelligence, Care Quality Commission
104. Ms Helen Street, Cambridge University Hospitals NHS Foundation Trust
105. Prof. Iain Buchan, Executive Dean; Chair in Public Health and Clinical Informatics, University of Liverpool
106. Dr. Iain Thomas, Cambridge Enterprise
107. Prof Sir. Ian Diamond, Office for National Statistics
108. Dr. Ian Dunham EMBL-EBI, Wellcome Sanger Institute and Open Targets
109. Ian Townend, Chief Architect, NHSX
110. Dr. Inesa Thomsen, Department of Health and Social Care
111. Dr. Ingrid Wolfe, Department of Women and Children's Health, School of Life Course Sciences, Faculty of Life Sciences and Medicine, King's College London
112. Prof. Isabel Oliver, Public health England
113. Jackie Gray, Executive Director, Privacy, Transparency & Ethics, NHS Digital
114. James Brook, U.K. and Ireland Head of Clinical Research, IQVIA
115. Prof James Carpenter, UCL & LSHTM
116. Mr James Holland, Methods Analytics
117. Prof. James Medcalf, UK Renal Association
118. Lord James O'Shaughnessy, House of Lords
119. Dr. James Squires, Academy of Medical Sciences
120. James Weatherall, Vice President, Data Science & Artificial Intelligence, R&D, AstraZeneca
121. Prof. Jane Sandall, King's College London
122. Jason du Preez, CEO, Privitar

123. Jaspreet Takhar, Baker McKenzie
124. Dr. Jeni Tennison, Open Data Institute
125. Dr. Jenni Burton, University of Glasgow
126. Prof. Jenny Hewison, Leeds Institute of Health Sciences, University of Leeds
127. Mrs Jenny Thomas, DigitalHealth.London
128. Jerry Clough, Vice President, Population Health Management, Optum UK
129. Jillian Hastings Ward, Participant Panel at Genomics England
130. Prof. Jim Davies, University of Oxford
131. Joanna Davinson. Central Digital and Data Office
132. Joe Edwards, The Association of the British Pharmaceutical Industry
133. Prof. John Appleby, The Nuffield Trust
134. John Bates, Senior Statistical Analyst, Department of Health and Social Care
135. Prof. John Bradley, Cambridge University Hospitals
136. Dr. John D Halamka, President of Platform, Mayo Clinic
137. Dr. John Parry, TPP
138. Mr. John Taysom, CSaP fellow, The University of Cambridge
139. Prof. Jonathan Kay, Faculty of Clinical Informatics
140. Prof Sir. Jonathan Montgomery, University College London
141. Dr. Jonathan Stokes, University of Manchester
142. Prof. Julia Hippisley-Cox, University of Oxford
143. Mrs Julia M Snowball, University of Bath
144. Ms Juliet Tizzard, Health Research Authority
145. Prof. Justin Keen, University of Leeds
146. 1Dr. Karen Kirkham, NHSE/I
147. Dr. Karen Laura Mansfield, University of Oxford Department of Psychiatry
148. Kate Cheema, Director of Health Intelligence, British Heart Foundation
149. Dr. Katherine Morley, RAND Europe
150. Dr. Katie Harron, UCL Great Ormond Street Institute of Child Health
151. Mrs Katie Tucker, Guy's and St Thomas' NHS Foundation Trust & Institute of Women and Children's Health
152. Dr. Katya Masconi-Yule, DigitalHealth. London
153. Keith Ridge, Chief Pharmaceutical Officer for England, NHS England and NHS Improvement
154. Mr Ken W Dunn, National Casemix Office, NHS Digital
155. Kenji Takeda, Director of Academic Health and AI Partnerships, Microsoft Research
156. Miss Kerrie Woods, Oxford University Hospitals NHS Foundation Trust
157. Prof. Kieran Walshe, Health and Care Research Wales
158. Kirsty Irvine, Independent Group Advising (NHS Digital) on Release of Data (IGARD)
159. Laura Bickerdike, Senior Policy Advisor, Office for Life Sciences
160. Dr. Laura Gilbert, 10 Downing Street
161. Ms Laura Wade-Gery, NHS Digital
162. Laurie Hawkins, CEO, AITIA Global

163. Mr Lea Milligan, MQ Mental Health Research
164. Léa Quentin, Workstream Lead, Kent Surrey Sussex Academic Health Science Network
165. Mr Lee Coulson, NHS Devon Clinical Commissioning Group
166. Prof. Leslie Mayhew, Ilc uk and the business school city university
167. Lisa Annaly, Head of Provider Analytics (Hospitals), CQC
168. Dr. Lisa Gibbons, South West Peninsula Clinical Research Network & Claremont Medical Practice
169. Louis Mosley, Palantir Technologies UK
170. Dr. Louise Wood, Director Science, Research & Evidence, Department of Health and Social Care
171. Lucy Vickers, Department of Health and Social Care
172. Mr Luke Readman, NHS London
173. Dr. Macey L Murray, MRC Clinical Trials Unit at University College London
174. Maddy Phipps-Taylor, CEO, Eva Health Technologies
175. Prof. Mahesh KB Parmar, University College London
176. Mallory Durran, Cabinet Office
177. Dr. Marcus Baw, Independent Clinical Informatician, General Practitioner, and Software Developer
178. Marcus Grazette, Europe Policy Lead, Privitar
179. Dr. Margaret O'Hara, Long Covid Support
180. Marie-Anne Demestihas, Senior Workstream Lead, Kent Surrey Sussex Academic Health Science Network
181. Dr. Mark A. Green, University of Liverpool
182. Mr Mark Deacon, The Brain Tumour Charity
183. Dr Mark Shepherd, Tencastle
184. Prof. Mark Thomson, STFC-UKRI
185. Dr. Mark Toal, Deputy Director of Research Systems, Department of Health and Social Care
186. Mr Markus Bolton, (Joint CEO), System C and Graphnet Care Alliance
187. Prof. Martin Landray, Nuffield Department of Population Health, University of Oxford
188. Prof. Martin O'Flaherty, University of Liverpool
189. Miss Mary Gough, The Company Chemists' Association
190. Matt Westmore, Chief Executive, Health Research Authority
191. Prof. Matthew R Sydes, MRC Clinical Trials Unit at UCL
192. Mr Matthew Swindells, MJS Healthcare Consulting
193. Ms Maxine Kennedy, NVIDIA
194. Melanie Leis, Institute of Global Health Innovation, Imperial College London
195. Menelas Nicolas Pangalos, EVP and President BioPharmaceuticals R&D, AstraZeneca
196. Dr. Merlin Dunlop, Ardens
197. Dr. Michael Chapman, NHS Digital
198. Mr Michael Cousins, Roche
199. Prof. Mireille Toledano, Imperial College London
200. Prof. Mohammed A Mohammed, Strategy Unit & University of Bradford

201. Mrs Rosemary Boyle, Senior commercial lawyer, University of Cambridge
202. Dr. Natalie Banner, Wellcome
203. Dr. Natalie Bohm, Pfizer Ltd
204. Mr Neil Mason, Methods Analytics
205. Neil Smart, Chair, Scottish Health Technologies Council, Scottish Health Technologies Group, HIS
206. Dr. Nicola Byrne, The National Data Guardian for Health and Social Care (NDG)
207. Dr. Nicole Mather, IBM
208. Nora Cooke O'Dowd, Head of Research and Intelligence, HFEA
209. Dr. Olly Butters, University of Liverpool
210. Lord Patrick Carter, NHS Improvement
211. Prof. Patrick Chinnery, Medical Research Council & UKRI
212. Prof. Paul Aylin, Imperial College London
213. Mr Paul Connell, ODI Leeds
214. Prof. Paul Elliott, Imperial College London
215. Prof. Paul Taylor, UCL
216. Prof. PB Jones, University of Cambridge
217. Dr. Pearse Keane, Moorfields Eye Hospital NHS Foundation Trust
218. Mr Pete Stokes, Office for National Statistics
219. Prof. Peter Bower, NIHR Clinical Research Network
220. Peter Bradley, Director of Health Intelligence/Chief Information Officer, Public Health England
221. Prof. Peter Denton White, Queen Mary University of London
222. Peter Spilsbury, Director,Strategy Unit, MLCSU
223. Petros Kotsidis, Chief Data Officer, FITFILE
224. Phil Earl, Deputy Director - Data Strategy, Implementation and Evidence, Department Digital, Culture, Media and Sport
225. Dr. Philip Quinlan, University of Nottingham
226. Philip Russmeyer, Founder and CEO FITFILE Group Limited
227. Dr Pia Hardelid, UCL
228. Polly Sinclair, Health Innovation Network
229. Ms Rachael Brannan, Public Health England
230. Rachael Graham, Sense
231. Dr. Rachel Charlton, University of Bath
232. Rachel Habergham, NHS Digital
233. Dr. Richard J.Q. McNally, Newcastle University
234. Mr Richard Jarvis, EMIS
235. Mr Richard Laux, Cabinet Office
236. Prof. Richard Wakeford, The University of Manchester
237. Richard Welpton, Head of Data Services Infrastructure, Economic and Social Research Council
238. Prof. Robert Bristow, Manchester Cancer Research Centre and University of Manchester
239. Dr. Robert McCombe, Information Commissioner's Office
240. Prof. Robert Stewart, King's College London
241. Robin Flaig, Deputy Director UK Longitudinal Linkage Collaboration, University of Edinburgh
242. Robin Sergeant, Chief Executive Officer, Optum UK
243. Rohan Allen, Analyst, Disability Unit
244. Ms Rose Dewey, NHS Bradford District and Craven CCG
245. Ms Rosie Richards, NHS Confederation
246. Prof. Ruth Gilbert, University College London
247. Sally Cavanagh, Clinical Information Manager, National Specialised Commissioning Business Informatics and Intelligence Team, NHS England and NHS Improvement
248. Sally McManus, Violence and Society Centre, City, University of London
249. Mrs Sam Organ, Public Health England
250. Samantha Jones, Prime Ministers Expert Advisor - NHS Transformation & Social Care, No 10 Downing Street
251. Sara Ward, Chief Operating Officer, Oxford Academic Health Partners (NIHR AHSC)
252. Prof. Sarah Elizabeth Rodgers, University of Liverpool
253. Ms Sarah Gold, Projects by IF
254. Sarah Scobie, Deputy Director of Research, Nuffield Trust
255. Mrs Sarah Shenow, MQ: Mental Health Research
256. Sarah Wilkinson, previously CEO, NHS Digital
257. Dr. Sarion Bowers, Wellcome Sanger Institute
258. Seb Bacon, CTO, The Bennett Institute for Applied Data Science, University of Oxford
259. Shane Tickell, CEO – Founder, Temple Black - Quantum Health Technologies
260. Prof. Sharon Love, MRC Clinical Trials Unit at UCL
261. Dr Simon Kolstoe, University of Portsmouth
262. Simon McDougall, Deputy Commissioner, Information Commissioner's Office
263. Dr. Simon Overell, Human.ai
264. Mr Simon Phipps, Meshed Insights Ltd
265. Dr. Sinead Savage, Manchester Cancer Research Centre, University of Manchester
266. Stacy Gorelik, Flatiron Health
267. Prof. Stephen W. Duffy, Queen versity of London
268. Dr. Steve Harris, University College London Hospital
269. Prof. Sue Pavitt, University of Leeds; NIHR CRN
270. Ms Susan Lyon, UK Renal Association
271. Dr. Susheel Varma, Health Data Research UK
272. Ms Tamsin Berry, Population Health Partners
273. Prof. Tamsin Jane Ford,
274. University of Cambridge
275. Mrs Tamsin Morris, AstraZeneca UK
276. Dr. Tanya Bleiker, President, British Association of Dermatologists
277. Sir Terence Stephenson
278. Chair, Health Research Authority
279. Dr. Thomas Anthony Ward, CQC
280. Tim Benson, R-Outcomes Ltd
281. Tim Coote, Chief Technical Officer, Anthropos Digital Care Ltd
282. Tim Donohoe, Director of Delivery, Assurance and Operations, NHSX
283. Prof. Tim Hubbard, King's College London

284. Dr. Tim Jobson, Taunton and Somerset NHS FT

285. Tina Woods, CEO, Longevity International

286. Tom Palser, Clinical Informatics Lead, Methods Analytics

287. Dr. Tony Calland, Chair Confidentiality Advice Group, Health Research Authority

288. Prof. Usha Menon, University College London

289. Vasilis Kapsalis, Senior Director, Commercial HPC and AI, DataDirect Networks

290. Vivienne Parry, Head of Public Engagement Genomics England

291. Wesam Baker, Director of Strategic Analytics, Economics and Population Health Management Mersey Care NHS Foundation Trust

292. Dr.William van't Hoff, NIHR Clinical Research Network

Groups

Thank you to the following groups who submitted written feedback on behalf of their organisations:

- medConfidential
- Data and policy experts from GSK
- Committee on Medical Aspects of Radiation in the Environment (COMARE)
- Our Future Health
- Karen Dennison Centre for Longitudinal Studies, University College London Social Research Institute
- MQ: Mental Health Data Science group
- MQ: Mental Health Research

Open focus groups

Thank you to all of the many participants (more than 170 in total) in the following eight single sector focus groups, and all those who registered their interest:

- Data ethicists
- Health and Care professional bodies
- Operational researchers
- SMEs
- NIHR researchers
- Expert patient and public representatives
- NHS Analysts
- Medical Research Charities

NHS and DHSC Policy Team

Thank you to the policy and communications team at both DHSC and NHSX who enabled this review to reach publication stage, this would not have been possible without their tireless efforts and support:

- Nicky Brassington - Deputy Director of Data & Analytics, NHSX
- Rebecca Fitton, NHSX
- Joseph Watts, NHSX
- Anna Steere, NHSX


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