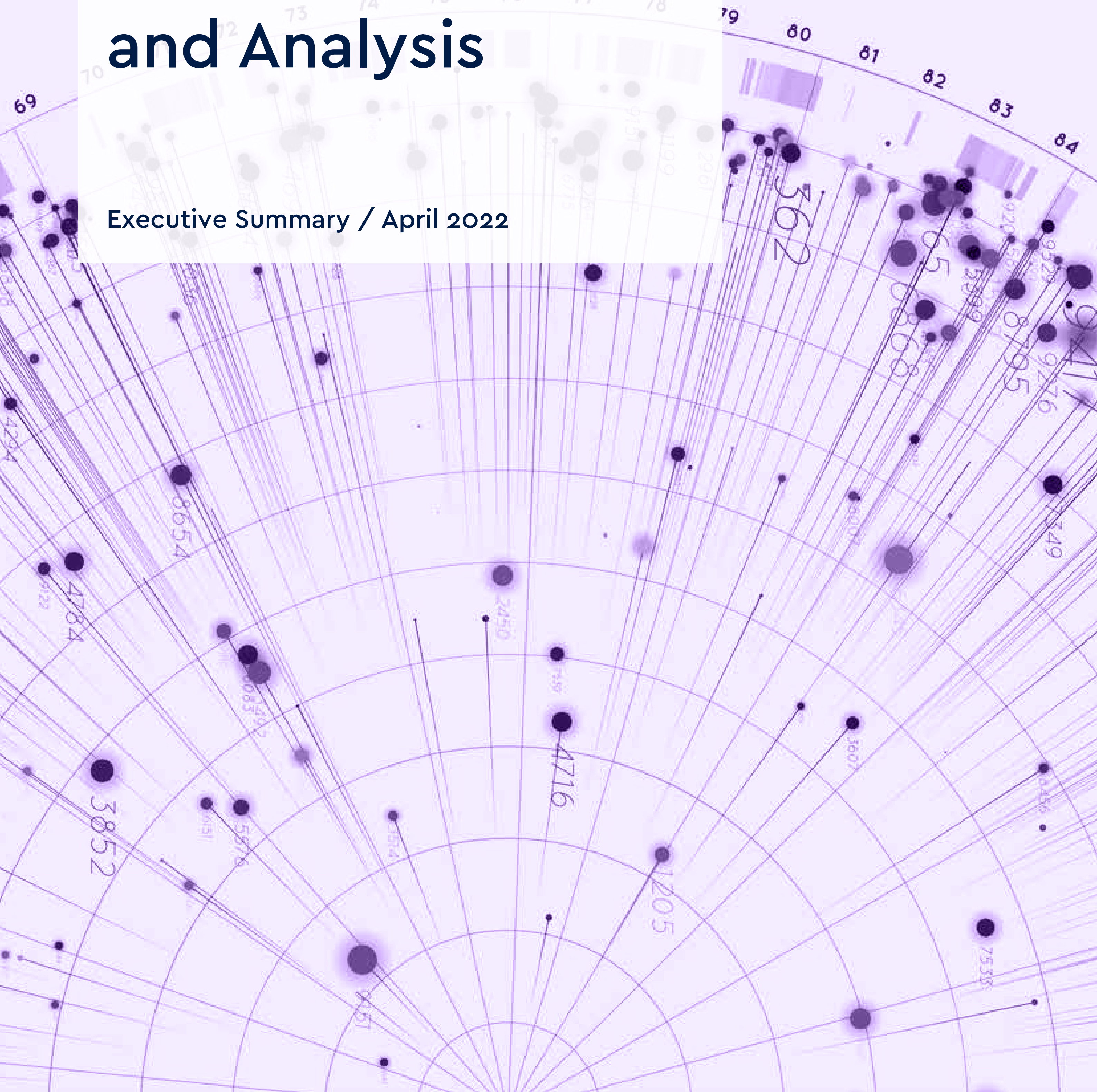
A review commissioned by the Secretary of State for Health and Social Care

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

Executive Summary / April 2022



Executive 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 full Terms of Reference are available in Appendix 1. The recommendations are derived from extensive engagement with over 200 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.
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.
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.
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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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

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 PPIE is done.

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.

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.

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.

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.

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.
    - Develop common working practices for the "generic compute and database layer" of TREs with generic skilled technical teams from private and public sectors.
    - 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.
    - 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.
    - Ensure all TRE teams work in the open, sharing and documenting all code and working methods as they go, to support adaptive innovation.
    - 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).

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.

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.

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