Health Data in the UK:
How we got here and where to go

William Bolton, Anmol Arora, Zoya Yasmine

Foreword

The UK possesses uniquely valuable health data assets, but the systems through which they are accessed, governed and used for research remain fragmented. Recent years have shown both the genuine public benefits that can arise from responsible data use and the difficulties that follow when governance, communication and transparency fall short. As national infrastructure evolves and artificial intelligence (AI) expands what can be done with health data, the challenge is no longer only technical. The most important questions ahead are largely social, ethical and political, and should focus on how to secure public trust in the use of health data.

How we got here

The NHS is used by almost all of the UK's population. As healthcare has become more technologically enabled, the data collected represents a rich source of information on individual and population-level health. Yet, within the NHS, no single entity holds all patients' data: individual GP practices, NHS Trusts, regional Integrated Care Systems (ICS), and social care providers can all be data controllers depending on the context. As the UK healthcare landscape increasingly begins to straddle both the public and private sectors, private clinics also collect and control their own patients' health data, while pharmaceutical companies generate new patient data through clinical trials. In addition, as data-driven medical technologies, wearables, and consumer testing services are becoming more commonly used, private technology corporations are also becoming holders of large patient datasets, which can reveal intricate and scientifically useful insights about individuals' health. Connecting an individual's health data between providers or even within the same system is often not trivial; for example, linking GP data with hospital-level data can be challenging. Taken together, over the last few decades, the UK's health data has been rapidly expanding across a distributed, fragmented and expanding network of public and private actors.

A disconnected and siloed health data landscape puts up barriers. It makes it difficult for researchers and policymakers to understand how and where to access the right data for health research. There have been several initiatives in the past that aimed to address or navigate the UK's fragmented health data system. However, for the most part, these have ultimately been abandoned or have not reached their desired potential, primarily because of the erosion in patient trust. The first example is *care.data* in 2014, which was an attempt to centralise GP records. Yet a failure to transparently communicate to patients how their data would be used and shared led to this project ending before any centralisation could be done

[1]. Following this attempt, the General Practice Data for Planning and Research (GPDPR) in 2021 aimed to extract GP patient data into a central database for research and planning, but received negative public responses. Notably, 1.2 million patients opted out of the initiative in a single month due to fears that the data would be sold to private firms [2]. Such hesitancy may have been derived from the fallout from the NHS Royal Free sharing 1.6 million patient records with Google DeepMind in 2015 for the development of medical software, Streams. The UK's data protection regulator, the Information Commissioner's Office, in this case, concluded that the data was shared unlawfully by the Royal Free [3].

Despite these examples of data sharing initiatives where legal and ethical principles have been infringed, there are also successes, though these are often less well known by the public. The Data Access Request Service has proven a useful mechanism to access NHS England data and services for approved purposes [4]. Genomics England started in 2013, under what is now named the Department of Health and Social Care, with the 100,000 Genomes Project [5]. It operates within its own Secure Data Environment (SDE) (also known as Trusted Research Environments [TREs]) and contains whole-genome sequencing data linked to limited clinical data. Similarly, UK Biobank, which started in 2003 and operates as a registered charity, is often noted as the world's most detailed, publicly accessible, large-scale health dataset. It is a long-term, prospective study to understand what influences disease, and so contains a wide variety of health data from the 500,000 volunteers [6]. As of early 2026, UK Biobank is now integrating coded anonymised GP records from across England into its existing dataset, including diagnoses, prescriptions, and referrals, but no free text or notes [7]. Additionally, CPRD and QResearch operate as not-for-profits and contain pseudonymised primary care data for patients registered at participating GP practices [8,9]. These examples show that health data sharing initiatives are possible, and that the especially sensitive nature of the data held by Genomics England and UK Biobank helped drive the creation of bespoke governance mechanisms which built trust.

The COVID-19 pandemic posed extraordinary circumstances for the health care system in the UK. To help tackle the pandemic, the Secretary of State for Health and Social Care issued Control of Patient Information (COPI) notices, which accelerated access and use of confidential patient information for purposes of pandemic response [10]. These public health legislative instruments could be used by researchers and sparked unprecedented data sharing and caused an increased investment in infrastructure like SDEs, which enable approved users to access pseudonymised data, run code and extract certain outputs. OpenSafely is one example of an SDE, born out of the pandemic, which has led to over 200+ projects, 100+ published outputs with 30+ organisations [11,12]. Other initiatives developed include the COVID-19 Data Store and Dashboard, which was developed in collaboration with commercial partners for tracking resources and the spread of the disease [13], and the RECOVERY trial, which used NHS data to recruit patients into the world's largest COVID treatment trial [14]. The circumstances of the pandemic, therefore, lowered barriers to accessing data and led to a huge increase in research output, demonstrating the potential benefits of increased health data sharing in controlled environments. However, since the pandemic ended, there has been some concern that some of these provisions have been overstretched and gone beyond the scope of pandemic response. Foresight is one example; this was a technically impressive AI model that could predict a patient's likely future health events from their longitudinal patient records [15,16]. This research project demonstrated the potential of large-scale AI systems to benefit public health. However, work

on it was paused in 2025, due to a BMA referral triggering an Information Commissioner's Office investigation [17]. Signalling the need for greater clarity on how large-scale AI research should be governed within NHS data environments. COVID-19 resulted in expanded health-data access, which enabled work that delivered substantial public benefit, but it has also exposed the difficulty of sustaining this public perception and culture once emergency conditions have passed.

In a broader push to modernise NHS data infrastructure, NHS England recently procured a Federated Data Platform (FDP) to connect data across NHS organisations to support operational decision-making and improve patient care. This was awarded to a consortium led by Palantir in November 2023 [18]. Similar to previous initiatives, the FDP has been subject to criticism and resistance from patients and clinicians. Following a similar line as the Royal Free-DeepMind collaboration, the FDP is being legally challenged by Fox Glove, Just Treatment, the Doctors' Association UK and the National Pensioners Convention. The campaigners have sent a "letter before claim" to the UK Government for an explanation of how the FDP will comply with the law [19]. Furthermore, individual NHS Trusts have hesitated to integrate Palantir's technology into their own infrastructure, such as Leeds Teaching Hospitals NHS Trust and Berkshire Healthcare NHS Foundation Trust [20]. Efforts to join up NHS data at scale are likely to keep encountering resistance unless they are matched by a clear legal basis, visible public-interest benefits, and proportionate and transparent governance arrangements capable of commanding trust from both patients and clinicians. The FDP sits within a broader recent movement towards more connected NHS data infrastructure, but has also continued the debates over how such systems should be governed and justified.

Lessons from the past

The history of health data in the UK offers several lessons and highlights why certain issues persist today. The fragmented nature of health data has been a root cause of many inefficiencies and downstream problems. For example, when setting up SDEs during the COVID-19 pandemic, many individual sites repeatedly duplicated efforts, including learning about and investing in infrastructure or cleaning data. The independence of various sites has led to numerous SDEs existing at differing levels around the UK, each with its own data quality, biases and linkage issues, as well as differing legal and governance requirements for research. This is not easy for researchers to navigate, and alongside long and opaque review processes, disincentivises new entrants and inter-trust research. Together, these aspects have drained sites' resources and delayed impactful innovation from occurring on a large scale. History shows us that significant value can be silently lost from the NHS and the UK by not effectively leveraging our rich data assets.

A major challenge in recent years has been the transition from the exceptional conditions of the COVID era to stable and repeatable pathways for routine research and innovation. Emergency permissions allowed the UK to move quickly during the pandemic, but they are not a stable basis for long-term research and innovation. When the legal basis for research varies across sites, entities, and over time, researchers must spend significant effort to navigate the often unclear landscape and are discouraged from pursuing impactful work that could benefit the NHS as a whole. The speed of innovation has decelerated since the

pandemic, with a justified return to a less permissive balance between speed, risk and benefit. However, a goal should be to establish streamlined pathways for access, analysis, evaluation and deployment that are lawful, legible and repeatable in normal times. This is especially important for AI research and development, where the rate of development and scale of both opportunity and risk are greater than in many traditional analytical settings.

SDEs are not generally designed for the training or evaluation needs of modern AI models. GPUs are an essential piece of equipment for this area of research, but they rarely exist for research within the NHS, leading to sites becoming dependent on and procuring these resources from big technology providers. This reliance on hyperscalers likely sits uneasily with the public expectations of NHS stewardship, sovereignty and control, and given that commercial terms are frequently opaque, this can undermine trust. Such firms can nevertheless bring genuine advantages in security, resilience and technical capability. Justifying high up-front expenditures, such as equipment or expensive technical talent within a cash-strapped NHS, is obviously challenging, leading to these problems often being kicked down the road. The complexity of this type of research can also often require some form of non-scalable and costly bespoke work, such as data labelling or linking. Given that the level of financial investment varies by site, those with more resources are naturally more attractive to researchers, leading to uneven data distributions, which can introduce biases that risk being perpetuated in research outputs. The problem to solve for AI research is therefore not purely technical, but rather how to securely set up and govern such work in a way that preserves public confidence, avoids lock-in, and secures fair long-term value for the health system.

Another persistent problem is the lack of a clear bridge between research and operations. At present, the boundary between research, service evaluation, service improvement and direct-care deployment is often blurred, especially for AI systems. SDEs often do not perfectly replicate real clinical scenarios, making it difficult to invest in and assess the potential of early-stage innovations. The lack of sandboxes and reliable pathways to translating findings has led to lots of research and technology being lost in the ‘valley of death’. Ambiguity in the translation process creates confusion, inconsistent approvals and the risk of trust-breaking failures when something developed in one context is used in another without a clearly justified transition. This problem does not only slow innovation, but also increases the likelihood that valuable work is lost or that deployment occurs without the appropriate safeguards.

Underlying everything to do with health data is a shortage of trust. Years of high-profile failures have made the public understandably wary of how their health data is used and if the authorities are able to appropriately manage it. National Data Opts-Outs have historically been subject to scrutiny and have an unclear scope, but concerns are now heightened given the perception of AI models' capabilities and their need for large amounts of personal data. History has taught us that good intentions alone are not enough. Without a deliberate and transparent effort to rebuild this trust, even well-designed national approaches risk being undermined by public opposition before they can deliver meaningful benefits.

Where we are today

The landscape in the UK is beginning to shift. Meaningful investment and structural efforts are being deployed to address the problem of a siloed and fragmented environment, with each SDE having its own rules, processes, and stipulations. Most significantly, the UK government last year announced a £600 million Health Data Research Service (HDRS), developed in partnership with Wellcome [21,22]. The HDRS aims to provide a single, secure gateway to access NHS SDEs through federated technology, supported by a coordinating layer for governance, standards, and oversight. By forming an independent organisation with a governing committee, a model that has worked well in the past with initiatives such as UK Biobank, the aim is to ensure accountability and control. Even so, the reliance on a single commercial provider means that vendor lock-in and loss of control are risks that require active management as the service matures.

Other recent developments include the Health Data Research Innovation Gateway, which provides a centralised catalogue of datasets and access routes, making it easier for researchers to identify and request relevant resources [23]. The Five Safes framework is standardising how SDEs operate and connect [24], and the pan-UK Data Governance Steering Group is providing generalisable policies, contracts, open data formats and technical standards for research on health data [25]. A Value Sharing Framework was also published in 2023 to ensure the NHS gets a fair share of any value arising from the use of its data [26]. Taken together, these initiatives represent a more coordinated approach to health data than the UK has previously achieved.

The regulatory and governance landscape for researchers continues to be complex. It spans CAG and Section 251 approvals, GDPR legal bases, ethics review, and NHS information security standards [27–29]. Recent changes to the Data Use and Access Act 2025 note that commercial scientific research can be considered under the scientific research exemption to the prohibition on processing special category data [30]. This is a clear signal to researchers and the life sciences industry of the UK's pro-innovation approach and could drive increased external investment in the future. Navigating the full stack remains a structural disadvantage, though, for those without a dedicated legal and compliance resource. Reducing this friction for legitimate research without weakening the protections that public trust depends on is one of the central challenges for the HDRS.

A future direction

In a stronger future state, the current fragmentation of SDEs would be reduced through more standardised approval, accreditation and governance processes that are understandable to researchers, delivery organisations and the public. HDRS is a promising step in that direction, but it should be understood as part of a wider ecosystem and interface effectively with other UK capabilities. This includes the NHS Research Secure Data Environment

Network, NHS England Secure Data Environment, and ONS Secure Research Service. It should also learn from related initiatives in Wales, Scotland, and internationally. The SAIL Databank, Research Data Scotland, Findata, OpenSAFELY, UK Biobank, and the emerging European Health Data Space all offer useful examples of governance, interoperability, access design, and value capture.

What is notably absent from the current HDRS discussion is a coherent treatment of AI. Questions about what AI use is permissible, what safeguards are required, what computing resources are needed, and how model training within these environments should be governed remain largely unanswered. Issues such as model training permissions, the handling and possible export of model weights, privacy risks from memorisation, requirements for logging and audit, bias assessment, red-teaming, and the governance of collaborations with external partners, and extraction of any commercial value all need answering. Other bodies in this space have been proactive. The MHRA, responsible for regulating medicines and medical devices in the UK, has issued guidance on AI as a software medical device and engaged with synthetic data approaches in collaboration with the PHG Foundation [31,32]. Closing this gap and providing clarity on the taxonomy of AI use cases and explicit rules for what is in scope, for example, by developing an HDRS AI strategy, should be a near-term priority.

Alongside federated learning, which is being used by HDRS, synthetic data could play an important enabling role. A national synthetic data sandbox, linked both to HDRS research workflows and to later real-world deployment routes, could allow developers to test code, demonstrate technical functionality, and resolve basic engineering issues before touching real patient data. Used properly, this could reduce unnecessary exposure to personal data, lower barriers to entry, and shorten the path from concept to formal evaluation. It is important to note, however, that synthetic data is not a substitute for governed access to real patient data. Perhaps, a sensible future model may be one in which applicants are expected to demonstrate basic feasibility in synthetic data before requesting access to real patient data via an SDE. Over time, privacy-enhancing technologies should accelerate wider cross-sector and cross-industry collaborations without increasing the risk of compromising personal data.

A more effective future system would not only streamline access with gateways but also make movement through that system more consistent and intelligible. One practical extension to access and approval would be a national secure-data “passport”. Under such a model, a researcher who has already completed recognised training, identity verification, and core governance checks could move between approved environments without repeating full onboarding each time, while still complying with local, dataset-specific and project-specific controls where required. Regarding translation, a better system would define the categories and stages of deployment more explicitly and establish visible approval routes for each. Where tools move toward operational use, the transition should trigger proportionate requirements around information governance, clinical safety, auditability, bias assessment, procurement controls and, where relevant, regulatory compliance. Existing NHS guidance on AI information governance and digital clinical safety provides an initial basis for this more explicit transition model. By solving these structural barriers, the UK could reduce duplicated effort and shorten the path from research to impact.

Transparency must sit alongside technical capability. Hospitals and data custodians should be able to clearly describe what data they hold, how it is structured, what its limitations are, and whether it is linked to other datasets. Low-fidelity synthetic data may offer a relatively safe way for organisations to publish data that reflects the structure and format of underlying datasets without exposing personal information. This is being operationalised by University College London Hospitals NHS Foundation Trust and UK Biobank. Alongside showcasing the data itself, a public register of data-use projects, with clearly stated public-benefit rationales, would help make the system more visible. The public should be able to understand what data-use projects are happening, what benefits they are expected to deliver, and what safeguards are in place. Success stories should also be actively communicated to show how data use can produce meaningful benefits when governed well. There is also a strong case for giving patients and advocacy groups a clearer route to shape priorities by submitting problems or unmet needs they would like research to address, and by providing simple mechanisms for feedback on how data is being used. There is already excellent representation from public and patient members on review committees for health data access, and this should be considered standard practice. The Health Research Authority's existing publication of research summaries offers a useful baseline towards a model that provides broader visibility.

Long-term sustainability will depend in part on value capture. The NHS should not give away access to strategically important data assets without securing a fair return, but the model for doing so needs to be practical rather than punitive. There is a risk that companies may seek to develop and test their innovations on NHS data because of its streamlined access before shifting their focus to implementation abroad, to larger markets. The NHS should explore fair-value return mechanisms that create a credible expectation of UK benefit where innovations have been materially developed using NHS-linked data, whether through domestic piloting, preferential access terms, capability transfer, or other forms of public return.

A central playbook for commercial terms, transparency obligations, and external partnerships would help create consistency. There already exist reference points within the UK data infrastructure, including UK Biobank, CPRD and the ONS, which use forms of cost-recovery models with tiered access and pricing structures based on applicant type and complexity. That value should not be defined only in money. Partnerships could also contribute to capability building, workforce development, and infrastructure. Investment in training, including routes such as the NHS Graduate Data Science Programme and health data science internships, is particularly important if the NHS is to recruit and retain the technical talent needed to manage advanced data environments well. Without stronger career paths and institutional support, the NHS will continue to struggle to compete for the people required to sustain future innovation without reliance on external companies in the long-term. Several large technology companies already offer training to NHS staff through the NHS Futures platforms, and in the future, agreements with such companies could have a broader scope.

Finally, success should be measured against concrete outcomes rather than institutional aspiration alone. Once established, HDRS should be judged by whether it improves data quality, reduces median time to approval and time to usable access, increases user satisfaction, supports more research without corresponding growth in duplicated effort, and

lowers friction in activities such as linkage and environment access. It should also track indicators of legitimacy and safety, including public awareness, opt-out trends, security incidents, and the quality as well as quantity of outputs. For AI-related work, evaluation should extend beyond pilot volume to include evidence of performance across subgroups, auditability, and whether systems deliver measurable service or health benefit in practice. A regular public report, such as an annual State of Health Data Use, could provide an important mechanism for accountability and another avenue to build public trust.

References

[1] National Data Guardian. In pursuit of balance: unlocking the power of data whilst preserving public trust 2022. https://www.gov.uk/government/news/in-pursuit-of-balance-unlocking-the-power-of-data-whilst-preserving-public-trust.

[2] Legacy – Parliament 2019–24 2024. https://committees.parliament.uk/publications/45172/documents/223686/default/.

[3] Google DeepMind and class action lawsuit. Information Commissioner's Office 2022. https://ico.org.uk/for-the-public/ico-40/google-deepmind-and-class-action-lawsuit/.

[4] Data Access Request Service (DARS). NHS England Digital 2026. https://digital.nhs.uk/services/data-access-request-service-dars.

[5] About Us. Genomics England n.d. https://www.genomicsengland.co.uk/about-us.

[6] Our history. UK Biobank n.d. https://www.ukbiobank.ac.uk/about-us/our-history/.

[7] Access to our participants' GP data. UK Biobank n.d. https://www.ukbiobank.ac.uk/about-our-data/types-of-data/healthcare-records/access-to-our-participants-gp-data/.

[8] Clinical Practice Research Datalink (CPRD) n.d. https://ukhealthdata.org/members/clinical-practice-research-datalink-cprd/.

[9] QResearch n.d. https://www.qresearch.org/.

[10] Control of patient information (COPI) notice - NHS England Digital 2020. https://digital.nhs.uk/coronavirus/coronavirus-covid-19-response-information-governance-hub/control-of-patient-information-copi-notice.

[11] OpenSAFELY documentation n.d. https://docs.opensafely.org/.

[12] Andrews C, Schultze A, Curtis H, Hulme W, Tazare J, Evans S, et al. OpenSAFELY: Representativeness of electronic health record platform OpenSAFELY-TPP data compared to the population of England. Wellcome Open Research 2022;7:191. https://doi.org/10.12688/wellcomeopenres.18010.1.

[13] NHS England. NHS COVID-19 Data Store n.d. https://www.england.nhs.uk/contact-us/privacy-notice/how-we-use-your-information/covid-19-response/nhs-covid-19-data-store/.

[14] The RECOVERY Collaborative Group. Dexamethasone in Hospitalized Patients with Covid-19. New England Journal of Medicine 2021;384:693–704. https://doi.org/10.1056/nejmoa2021436.

[15] Kraljevic Z, Bean D, Shek A, Bendayan R, Hemingway H, Yeung JA, et al. Foresight-a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. The Lancet Digital Health 2024;6. https://doi.org/10.1016/S2589-7500(24)00025-6.

[16] Foresight AI case study - NHS England Digital 2025. https://digital.nhs.uk/data-and-information/research-powered-by-data/life-saving-research/case-studies/foresight-ai.

[17] Stephen Armstrong. NHS England faces investigation over granting Foresight access to GP patient data. BMJ 2025;389. https://doi.org/10.1136/bmj.r1192.

[18] NHS England. Contract explainer n.d. https://www.england.nhs.uk/digitaltechnology/nhs-federated-data-platform/security-privacy/contract-explainer/.

[19] Foxglove. Legal action launched: no legal basis for the £330 million Palantir NHS Federated Data Platform 2023. https://www.foxglove.org.uk/2023/11/30/legal-action-palantir-nhs-federated-data-platform/.

[20] Lindsay Clark. NHS England hospitals cast doubt on Palantir use case 2025. https://www.theregister.com/2025/05/16/nhs_hospitals_palantir/.

[21] National data service will simplify access to health data for research. Wellcome 2025. https://wellcome.org/insights/articles/national-data-service-will-simplify-access-health-data-research.

[22] Department of Health and Social Care. Prime Minister turbocharges medical research. GOVUK 2025. https://www.gov.uk/government/news/prime-minister-turbocharges-medical-research.

[23] Health Data Research Gateway n.d. https://healthdatagateway.org/en.

[24] Government Digital Service. The Five Safes Framework. GOVUK 2025. https://www.gov.uk/data-ethics-guidance/the-five-safes-framework.

[25] UK Health Data Research Alliance. Data Access and Governance n.d. https://ukhealthdata.org/trust-and-transparency/data-access-and-governance/.

[26] Value Sharing Framework for NHS data partnerships. NHS England Transformation Directorate 2023. https://transform.england.nhs.uk/key-tools-and-info/centre-improving-data-collaboration/value-sharing-framework-for-nhs-data-partnerships/.

[27] Confidential patient information and the regulations. Health Research Authority 2023. https://www.hra.nhs.uk/about-us/committees-and-services/confidentiality-advisory-group/confidential-patient-information-and-regulations/.

[28] Legal basis for processing data. Health Research Authority 2018. https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/data-protection-and-information-governance/gdpr-detailed-guidance/legal-basis-processing-data/.

[29] Data Security and Protection Toolkit n.d. https://www.dsptoolkit.nhs.uk/ (accessed March 13, 2026).

[30] Data (Use and Access) Act 2025: data protection and privacy changes. GOVUK 2025. https://www.gov.uk/guidance/data-use-and-access-act-2025-data-protection-and-privacy-changes.

[31] Software and artificial intelligence (AI) as a medical device. GOVUK 2025. https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device.

[32] PHG Foundation, Reich V, Mitchell C, Redrup Hill E, Myles P, Branson R. Synthetic data for development of AI as a medical device (AIaMDs): Regulatory considerations. PHG Foundation; 2025. https://doi.org/10.61599/vhac1.