Health Data Research Service (HDRS) Digital Ecosystem Analysis

Final report

wellcome

EMRYS HEALTH

nesta

Contents

Scope of this report 3

Executive summary 4

Introduction 6

Methodology 9

The current landscape 12

What capabilities are present? 28

What technologies are missing? 33

What are the constraints? 36

From gaps to opportunities 39

Pilot initiatives 52

Conclusion 55

Glossary 56

Acknowledgements and attributions 58

About the authors 60

References 61

Scope of this report

This review was commissioned by Wellcome as "an analysis of the digital technology supporting the UK Health Data ecosystem". It provides a comprehensive, evidence-based assessment of the technical and clinical digital infrastructure underpinning those assets; mapping what exists, what works, and where critical gaps constrain the UK's ability to compete globally in health research and life sciences.

This review examines the technology supporting the health systems of all four UK nations. It identifies what enables or constrains delivery of the six core Health Data Research Service (HDRS) capabilities and suggests where interventions might have the most impact.

The scope encompasses the full data flow from source systems through to the service provision layer accessed by researchers, covering both technical components and real-world legal, organisational and operational constraints.

The review does not:

- comprehensively catalogue all UK health data assets
- replicate existing stakeholder engagement already completed elsewhere
- analyse the interplay between the HDRS and other government schemes
- attempt to create a blueprint for the HDRS.

This review represents the independent findings of a specialist team and is not intended to represent the views or intentions of the HDRS team. It highlights potential opportunities to turn the UK's distributed assets into a coherent, high-performance foundation for data-driven discovery, that can help secure the UK's international reputation as a leading location for health research and innovation.

Executive summary

The UK health data ecosystem is at a critical juncture, and the HDRS presents a once-in-a-generation opportunity to harness data for the health and wealth of the nations of the UK.

It is apparent that there is potential for enormous public benefit from insights gained from routine treatment in one of the world's largest publicly-funded health systems. Similarly, access to data from the UK's diverse population could attract research and development (R&D) investment in the life sciences sector and speed the development of new treatments with national and global applications. However, the scale, richness and complexity of the data environment also present a challenge – the heterogeneity encountered in terms of data standards, infrastructure, governance and quality mean it can be a daunting and time-consuming task to link and analyse data on a population scale or to track individual patients through their treatment pathways.

All of the potential benefits from the HDRS depend on building and preserving trust with the public. Recent data breaches reveal the limits of contractual mechanisms to prevent actors who seek to exploit data for commercial gain. As a consequence an increased and necessary level of scrutiny is being placed on security and privacy. Technical controls that provide temporary access to secure environments should exist alongside a culture where the needs of researchers do not outweigh the responsibility of organisations to keep data safe.

This independent review provides a comprehensive assessment of the digital infrastructure underpinning the UK's health data assets – mapping what exists, what works, and where critical gaps constrain our collective ability to compete at the very highest level globally in health research and life sciences.

In line with objectives of the HDRS, the review focused on the unique technical considerations of scientific discovery and research, rather than healthcare operational needs. Findings were informed by: semi-structured interviews with stakeholders across 36 organisations, validation workshops, in-person sessions across the four nations, a review of published and grey literature sources, and written submissions from organisations spanning academia, NHS, industry, and government. Findings were tested against six concrete user stories representing communities the HDRS must serve.

It shows that the HDRS could act as a force for improving data value by building on proven capability, and a coordinating layer that provides a reliable, professionally operated, and accountable service to its users. Success will require sustained attention on standardisation and integration, creating true buy-in with stakeholders across all four nations of a kind not previously achieved in the UK.

Summary of challenges

The UK has a wealth of health data that is too often trapped in systems that are difficult to access, making it difficult for researchers to benefit from the full breadth of data 'available'. Researchers struggle to link data for an individual, join up data that is hosted in separate organisations or geographies, and find the compute needed to analyse complex data. In practice, this fragmentation slows down clinical trials and reduces global investment – the lack of a joined-up system makes it harder for hospitals to find and recruit the right patients as trial participants. Too much valuable time is wasted attempting (and often failing) to access our most valuable health data due to poorly linked infrastructure and unpredictable processes.

Non-programmatic interventions

This report identifies a number of interventions that an incoming HDRS team might consider when addressing these challenges. They are not intended to be exhaustive or directive.

Defining minimum information standards for what data assets should contain, based on researcher needs and how that data should be represented; and pairing this with investment into accessing high-value data types that are most frequently missing – including hospital prescribing data, laboratory and pathology results, and clinical information currently locked in unstructured text.

Building a UK-wide data integration layer to enable cross-asset, cross-region, and cross-nation research. Different architectural approaches (centralisation, federated analytics, or a metadata-driven fabric) each carry distinct trade-offs that should be evaluated against user requirements and implementation timelines.

Establishing national data linkage services with agreed methods and transparent quality metrics would address one of the most frequently cited bottlenecks in the review. Standardised linkage methods would de-risk project delivery and underpin comprehensive longitudinal records while published quality metrics would allow researchers to account for linkage error.

Introducing Trusted Research Environment (TRE) accreditation standards that address usability alongside security would create accountability for the environments through which users access HDRS data and enable credible service level commitments. Addressing adequate compute capacity co-located with the data is essential for AI and computationally intensive workloads.

Digitising the governance layer by incorporating portable researcher credentials, machine-readable agreement templates, and transparent project lifecycle tracking could replace fragmented, manual, and opaque access processes that undermine UK competitiveness. Researchers would apply once, track progress in a single environment, and carry recognised credentials across data controllers. Users and public would have guaranteed visibility into how NHS data is being used, and to what effect.

Alongside these interventions, the review identifies system-level enablers that are also dependencies for technical investments (e.g., financial sustainability through transitioning from grant-funded projects to a professional service model, establishing value-return frameworks, embedding transparency and public engagement, aligning procurement practices etc.).

Illustrative pilot interventions

To address existing ecosystem challenges, this report articulates six hypothetical pilots. These are designed to build on capabilities that are available today, with documented market demand, and a clear route to measurable impact. They are explicitly designed to be illustrative, not prescriptive.

Each could deliver demonstrable value to end-users within realistic timeframes, while also building reusable technical components as dependencies. (See page 52 for details).

1. UK-wide Real-World Evidence asset linking GP, hospital, and prescribing data
2. UK-scale biomarker-enriched pan-cancer cohort for clinical trials
3. UK population epidemiology asset with cross-sector and cross-border linkage
4. Sovereign AI foundation model with linked digital pathology
5. HDRS digital governance and transaction management platform
6. Near-real time device surveillance platform

Conclusion

For the UK public to benefit fully from the rich health datasets we hold, there is an urgent need to change the way the nations' health data is collected, stored and accessed. Moreover, without this the UK risks being able to fully realise the economic, research and health potential of advances in data science, AI and computation.

An improved UK health data research ecosystem would attract substantial investment, accelerate clinical research, and create tangible benefits for UK public health. The UK, with universal free healthcare provision and diverse population is in a global position of strength but action is needed to realise these assets.Strengthening and connecting existing infrastructure, rather than replacing it, would build a trusted, modern, distributed service. This service would invest in scaling the parts of the system that work and fixing the components that are broken; accelerating trials, attracting global investment, enabling scalable research, and ultimately improving patient outcomes.

The components to build HDRS already exist, the gaps are in the connective architecture and service infrastructure to bring them together. The opportunity for revolutionary change comes only rarely in any area of science, and it is clear that this moment has come for HDRS.

Introduction

The HDRS represents an important opportunity to harness the UK's health data for research, aligned with the government's ambitions for NHS transformation and economic growth set out in the AI Opportunities Action Plan[1], 10 Year Health Plan for England: Fit for the Future[2], and Life Sciences Sector Plan[3].

The UK possesses world-class data capabilities, including genomics, primary care data, specialist registries, and cohort studies, but these assets currently operate in isolation[4,5]. The HDRS can create the technical infrastructure, operational capacity and governance processes to join these capabilities into a coherent national system. This requires working pragmatically – building on decades of investment across institutions, regions and nations – whilst establishing the technical foundation and the governance for the UK to compete globally in life sciences and health research.

This report provides a comprehensive assessment of the digital infrastructure landscape. It serves as an independent evidence base to inform HDRS leadership on what exists across the four nations, their current strengths and the capability gaps between what exists today and what is needed for a successful HDRS.

The latter part of this report describes potential opportunities that address current capability gaps and could drive future HDRS value. The overarching aim is to empower the HDRS to move quickly: from inception to delivery of high-value national projects that deliver measurable benefit to the UK population.

Scope of analysis

This independent review examines the technology supporting the data ecosystems across all four UK nations – England, Scotland, Wales and Northern Ireland – identifying what enables or constrains delivery of the six core HDRS capabilities, and recommending where strategic interventions can yield the greatest impact.

The scope encompasses the full data flow from source systems where data originates, through processing pipelines and research data assets, to integration layers, TREs, and the service provision layer through which researchers access the system. It addresses both technical components and the real-world constraints (legal, organisational, and operational) within which technology must function.

This analysis deliberately excludes several areas that, whilst important, fall outside its primary remit. It does not produce a comprehensive catalogue of all UK health data assets (numerous such catalogues already exist, and replicating this work would add limited value). The review does not replicate existing engagement work or stakeholder consultations already completed by other programmes. It does not analyse the interplay between HDRS and other government schemes such as the Voluntary Pricing and Growth (VPAG) programme, recognising that these intersections require separate strategic consideration.

Importantly, this report does not aim to design the HDRS or provide detailed financial modelling beyond a high-level assessment of the commercial research value of potential initiatives.

Background

This section examines how the UK health data ecosystem reached its present state and the strategic context shaping infrastructure requirements.

Diversity of infrastructure and academic focus

The UK health data landscape has developed organically rather than through coordinated national planning. Over the past two decades, multiple reviews have diagnosed similar problems whilst successive investments have created pockets of excellence without resolving fundamental fragmentation.

The Wachter Review (2016) analysed the failure of the National Programme for IT and examined digital maturity in secondary care[6]. The Goldacre Review (2022)[5] made extensive recommendations for England on efficient and safe use of health data, including TREs and reproducible analytical pipelines. The Sudlow Review recognised the need for a UK-wide strategic approach to overcome bottlenecks and break down siloes[4]. However, with the notable exception of the O'Shaughnessy (2023)[7] review of clinical trials, reviews have maintained a focus on academic research rather than commercial research. This has shaped infrastructure investments. Initiatives such as NIHR Biomedical Research Centres (BRCs), the Administrative Data Research UK (ADR UK) network, and DARE UK have primarily funded academic groups to develop research infrastructure. These have understandably focused on developing the tools and systems that support academic research but have (with some notable exceptions) deprioritised integrations with live services, support for commercial researchers and clinical trial delivery.

Investments also span multiple programmes operating without systematic coordination: NHS England's Federated

Data Platform for operational analytics[8]; the Secure Data Environment network for regional research access[9]; Health Data Research UK's (HDR UK's) discovery tools and governance standards[10]; devolved nation infrastructure including SAIL (Wales), the Scottish Safe Haven Network, and Northern Ireland's Honest Broker Service; OpenSAFELY's approach to primary care data[11]; and large cohorts in UK Biobank[12], Genomics England[13], and now Our Future Health[14].

Each addresses legitimate needs, yet operates within its own governance structures, with distinct technical approaches, and serves specific user communities[5]. The Single Patient Record (SPR), which seeks to consolidate siloed records into a 'single version of the truth', offers a potential foundation by mandating national interoperability standards. However, it remains in early development with full access not expected until 2028[2]. With a focus on direct patient care, it does not provide an immediate solution to the structural fragmentation of the research ecosystem.

The UK has developed world class components for research infrastructure but not yet a coherent national capability. Building new infrastructure risks duplicating mature capabilities, disrupting established relationships, and consuming resources that could be better deployed scaling what already works. At the same time, the nature of health data research is changing with the advent of AI. It is, therefore, essential to understand not only what infrastructure currently exists but what will be required in the near future.

The convergence of research use cases

A critical shift informing infrastructure design is that traditional boundaries between research types have collapsed. Two decades ago, epidemiological research, clinical trials, real-world evidence generation, commissioning analytics, and AI development might have been treated as distinct activities requiring separate infrastructures. Today, these distinctions no longer reflect how research is actually conducted[15,16].

Modern clinical trials draw on data for feasibility and recruitment and incorporate real-world evidence as external control arms[17]. Population health surveillance and commissioning decisions draw on the same linked datasets[18]. Real-world evidence studies for regulatory submissions require the same level of data depth as pharmacovigilance research[19]. In general, the traditional separation between 'research' and 'operations', between 'academic' and 'commercial', and between 'trials' and 'observational studies' no longer maps cleanly to infrastructure requirements.

At the same time, the UK's ambition to be a leader in AI[1,20] creates additional infrastructure requirements that, except in bespoke instances, current systems do not meet. Training multimodal models requires datasets that are larger and more diverse than those used for conventional statistical analyses[21]. Validating algorithm performance across demographic groups requires representative data[22] that current volunteer cohorts cannot provide. Regulatory approval increasingly demands real-world evidence of safety and effectiveness. Infrastructure designed primarily for epidemiological research will not be sufficient to deliver AI-enabled healthcare.[23]

These developments have implications for HDRS design. It is the analytical workload that will determine the optimal technical architecture, not the scientific discipline. Infrastructure optimised for a single workload, or algorithmic technique, will constrain others. It is not a question of which infrastructure approach is best, it is a question of how we prioritise the research use cases based on their analytical workload and match them to the teams and technology that can deliver them.

Changing global landscape

The UK has world-class health data, but the global health research market is expanding and changing the extent to which this offers the UK a competitive advantage. Two notable international examples are the European Health Data Space (EHDS) and Canada's Connected Care for Canadians Act.

The EHDS aims to establish a 'single market' for health data, mandating interoperability and streamlined access for a population of 450 million[24,25]. As the EHDS matures, it may establish a high-performance benchmark for the secondary use of data, placing immense pressure on the UK to transition from a 'federation of fragments' into a professionalised, national service.

Canada's Connected Care for Canadians Act, aims to prevent 'data blocking' and unlock similar research opportunities as the HDRS by requiring vendors to share information in secure, digital formats and ensuring data follows the patient[26].

The UK must retain and develop its ability to support world-class research, innovation, and AI development, while driving health system efficiency through clear, connected and privacy-protecting rules for data use at scale. This is especially true for commercial research and clinical trials – domains where the UK is already facing strong competition.

Underserved opportunities: commercial research and clinical trials

The Sudlow Review (2024)[4] highlighted that UK data landscape fragmentation and uncertain access timelines compromises ability to deliver research studies at scale.

The O'Shaughnessy Review (2023)[7] identified the UK's clinical trials environment as slow and failing to capitalise on the NHS's inherent advantages. Industry stakeholders describe the UK as "an unreliable and unpredictable partner", reporting the UK as the second slowest of 18 European countries for trial setup. Performance data confirms these concerns.

Industry-sponsored trial enrolment fell to just over 19,000 participants in 2024/25 – a seven-year low[27]. The

government has established targets of 150 days for trial setup by March 2026, and quadrupling in participant recruitment by 2029[3, 28]. These are order-of-magnitude changes, but the economic case is substantial: £3 billion gross value added, £485 million in NHS revenue, and 26,000 jobs[27]. It is essential, therefore, that the HDRS supports these ambitions.

These challenges persist not because of a lack of capability or intent, but because no existing organisation is mandated to set cross-cutting standards, align access models, or manage interoperability across the system as a whole.

A way forward

The UK health data ecosystem comprises world-class components developed over decades, yet these operate largely in isolation. Stakeholder engagement across all four nations reveals consensus on three principles: strategic direction matters more than technological sophistication, attempting comprehensive transformation risks delivering nothing, and progress depends on building upon proven capability rather than replacement.

The challenges documented in this review – fragmented data assets, inconsistent technical standards, unpredictable governance processes, and misaligned incentives – cannot be resolved through additional funding alone, through local optimisation, or through voluntary coordination. What is required is deliberate infrastructure that connects existing strengths into a functioning national system.

This report examines what currently exists, identifies where capability is strong and where critical gaps constrain research delivery, and analyses the technological and non-technological factors that determine what the HDRS can achieve. The opportunities presented address specific, documented barriers. The pilot initiatives are illustrative examples of how targeted interventions can simultaneously deliver measurable value to end-users whilst building reusable technical components that become national assets.

Throughout this analysis, the HDRS is framed as a service rather than simply infrastructure. Service delivery requires reliability in access timelines, responsiveness to user needs, professional-grade technical capability, and recognition that researchers and commercial partners are customers whose requirements shape what must be built. These service principles inform both the gap analysis and the opportunities that follow.

Methodology

This section describes how evidence was gathered, analysed, and how recommendations were developed and tested.

Evidence gathering

Evidence was gathered through three primary channels: stakeholder interviews, literature review, and written submissions. These sources served complementary purposes. Interviews captured the current operational reality and practitioner perspectives. The literature provided technical details, international comparisons, and historical context. Written submissions enabled organisations to articulate strategic priorities in their own terms.

Stakeholder interviews

We conducted 39 semi-structured interviews with 78 stakeholders across 36 organisations, spanning: academic institutions, NHS organisations, government bodies, commercial data providers, life sciences companies, technology vendors, and charities across all four UK nations. Interviews followed a common structure but were adapted to each stakeholder's domain.

Interview topics included: current data assets and technical infrastructure, data flows from source systems to research environments, governance and access processes, user experience and pain points, non-technical constraints, commercial models and sustainability, and perceived gaps and development priorities.

Interviews with industry stakeholders also explored research considerations, including clinical trials, real-world evidence studies, and AI development, and compared them with international alternatives.

Workshops

We held two virtual stakeholder engagement workshops with around ~45 attendees at each from the four nations and three in-person workshops in Northern Ireland, Scotland and Wales.

The first virtual workshop was meant as a validation workshop to gather feedback from key stakeholders to validate, further nuance or challenge our findings, and to ensure we were not missing anything critical. The second workshop was focused on the co-development and refinement of potential pilot initiatives in the context of the technology archetypes and access mechanisms we have identified.

The in-person workshops held in Northern Ireland, Scotland and Wales were used to gain an in-depth understanding of the technological and data landscape across the devolved nations.

Patient and public involvement and engagement sessions (PPIE)

We held two online PPIE sessions in collaboration with South West Analytics and Infrastructure Group in Healthcare, who convened a diverse and informed group of eight PPIE members. Participants provided feedback on the review's main findings and challenges, and discussed a selection of pilot ideas to inform the landscape review's recommendations.

Academic and grey literature

We conducted a literature review of 248 pieces of published research on health data infrastructure, UK policy documents, technical documentation from major data assets, and international comparisons with health data systems in comparable nations. Academic sources were identified through searches of Google Scholar, PubMed, and Scopus. Grey literature included previous UK government reviews[4–7], DARE UK[29] infrastructure landscape assessments, technical documentation from major data assets (Clinical Practice Research Datalink (CPRD), UK Biobank, Genomics England, OpenSAFELY, SAIL), and published strategies from England,[2,30,31] devolved administrations, and HDR UK. International comparisons drew on documentation from comparable systems, including those in Denmark, Finland, Israel, and the US.

Written submissions

We invited and reviewed submissions from key stakeholders, detailing current capabilities, identified gaps, and strategic priorities for infrastructure development. Submissions were received from 22 organisations, including major public research institutes, academic institutions, NHS organisations and life sciences companies.

Synthesis and analysis

Evidence synthesis

Evidence synthesis was informed by realist principles, seeking to understand specifics of what works, for whom, and under what circumstances (rather than cataloguing assets or assessing capability in the abstract)[32,33].

To structure evidence and assess gaps systematically, information was extracted using the ITPOSMO framework[34], which categorises system components across seven dimensions:

- **Information:** Data types, sources, quality, and standards
- **Technology:** Software, hardware, networks, and technical architectures

Table 1. User stories

Access to comprehensive health records	*"I am a regulator. I need safety surveillance and real-world outcomes data capture with high population coverage, so I can rapidly identify safety signals and access datasets for comparative effectiveness analysis."*
Research-ready datasets	*"I am a precision medicine researcher. I need to link multiomics data from consented cohort participants with their full NHS longitudinal records, so I can identify disease subtypes, validate biomarkers, and support the design of targeted therapeutics."*
Opening up advanced diagnostics data	*"I am a medtech developing AI-based tools. I need access to representative and multimodal data on high-performance compute, with joined-up capabilities for testing my models in real clinical systems, so I can seamlessly move from early development to market."*
Faster clinical trials	*"I am a clinical trial sponsor. I need integrated tools for feasibility assessment, site selection, and patient identification and recruitment, so I can deliver trials faster and more cost-effectively."*
Simpler access	*"I am a pharma customer. I need rapid discovery, rapid feasibility, transparent pricing, and guaranteed 30-day access to conduct time-sensitive epidemiological studies."*
Linking data for greater impact	*"I am an academic researcher. I need longitudinal data linking health outcomes with social determinants and environmental factors, so I can understand disease patterns, measure care gaps, and identify ways to improve the health of our population."*

- **Processes:** Workflows, operational procedures, and service delivery mechanisms
- **Objectives and values:** Goals driving system development and ethical principles governing data use
- **Staffing and skills:** Workforce capability, expertise gaps, and capacity constraints
- **Management and governance:** Oversight structures, accountability mechanisms, and decision-making frameworks
- **Other resources:** Funding models, contractual arrangements, and material infrastructure

This framework recognises that successful infrastructure requires coherent functioning across all seven categories.

Information was triangulated to construct the landscape description. Interview accounts of operational challenges were validated against technical documentation and literature. Claims about capability were cross-referenced across multiple stakeholders. The resulting landscape picture represents a synthesis across sources rather than any single account.

User stories

To test findings against core HDRS capabilities, six high-value user stories were constructed, validated against contemporaneous Department of Health and Social Care (DHSC) customer research and market analysis. These represent the primary user communities the HDRS must serve: clinical trial sponsors, pharmaceutical companies, regulators, AI/medtech developers, academic researchers, and precision medicine researchers. Full user stories are provided in Table 1.

These user stories were used throughout the gap analysis to ground technological findings against the delivery of an effective service to end-users.

Gap analysis

For each of the six core HDRS capabilities, the analysis considered an ideal state – what 'good' would look like if the capability were fully realised. These descriptions combined user needs (the requirements of different research communities), functional specifications (the operations the system must perform to serve these needs), technical requirements (the components that must exist and how they interact), and real-world constraints (operational and governance factors).

The gap analysis systematically compared the current capability against the ideal state. This was not a binary assessment of presence or absence, but a nuanced evaluation recognising that capabilities may exist in some locations but not others, components may be technically feasible but operationally constrained, pockets of excellence may lack mechanisms to scale, and infrastructure may serve some user communities effectively whilst failing others.

The gap analysis identified where the current landscape falls short, but equally importantly, where existing capability works well and can be built upon.

Validation workshop

A validation workshop was conducted with data leaders, researchers, and industry representatives. This was attended by 38 participants. The review team presented preliminary results from the evidence synthesis and the gap analysis. Participants were tasked with challenging assumptions, identifying gaps in the evidence base, and validating the initial findings.

Developing and evaluating initiatives

Opportunities analysis

Gap analysis findings informed an opportunities analysis that explored different technological options. These are specifications for infrastructure, standards, and/or other technologies, in areas that are critical gaps in the current landscape. Each option had to meet three criteria, it must:

- reflect a critical gap identified in the analysis

- not replace existing mature capability
- be technically feasible to accomplish within realistic resource and governance constraints.

Following the evidence synthesis, candidate options were assessed against the user stories and evaluated by the review team against additional criteria. These included:

- whether an intervention would resolve a structural UK-wide barrier
- whether an intervention could be implemented within existing legal frameworks
- whether an intervention would contribute reusable components to the broader HDRS architecture.

Options that addressed only narrow use cases or required extensive new legislation were deprioritised in favour of those with broader applicability and faster implementation paths. It is worth noting however that it may be challenging to address all opportunities and launch all pilots simultaneously.

Pilot initiatives

Pilot initiatives were proposed that might deliver rapid impact for HDRS end-users whilst laying the foundations for longer-term transformation. These initiatives are intended to be illustrative examples of what HDRS could deliver in its initial phase.

These were generated through three routes: direct derivation from gap analysis (where specific gaps could be crossed while delivering a pilot), direct stakeholder proposals (captured through interviews and workshops), and adaptation of existing initiatives that demonstrated partial capability requiring extension or integration. Each initiative was evaluated against selection criteria developed from the evidence base and refined through stakeholder input. These assessed whether an initiative could:

- through its execution, resolve structural barriers to UK research
- deliver demonstrable benefit within 12 months
- target verifiable market demand supporting HDRS financial sustainability
- build reusable technical components
- scale across the four nations
- deliver patient benefits that maintain social licence.

Initiatives were further assessed for feasibility based on governance friction (legislative or multi-controller requirements), technical complexity (data movement or new infrastructure requirements), and operational readiness (availability of existing teams versus need for new recruitment).

Pilot validation workshop

A second workshop focused on co-developing and evaluating candidate pilot initiatives, attended by 32 participants from 28 organisations. The review team presented pilots with supporting rationale. Participants assessed each pilot against the selection criteria, identified implementation risks, proposed modifications, and suggested alternative approaches. Workshop outputs directly informed the final pilot proposals.

Initial testing with patients and the public

Two initial workshops with patient representatives and members of the public explored attitudes toward health data use, priorities for system development, and concerns about privacy, commercialisation, and equity. These sessions tested the proposed capabilities, pilot selection criteria, and pilots against public expectations. Feedback from these sessions was incorporated directly into the findings of this report.

The current landscape

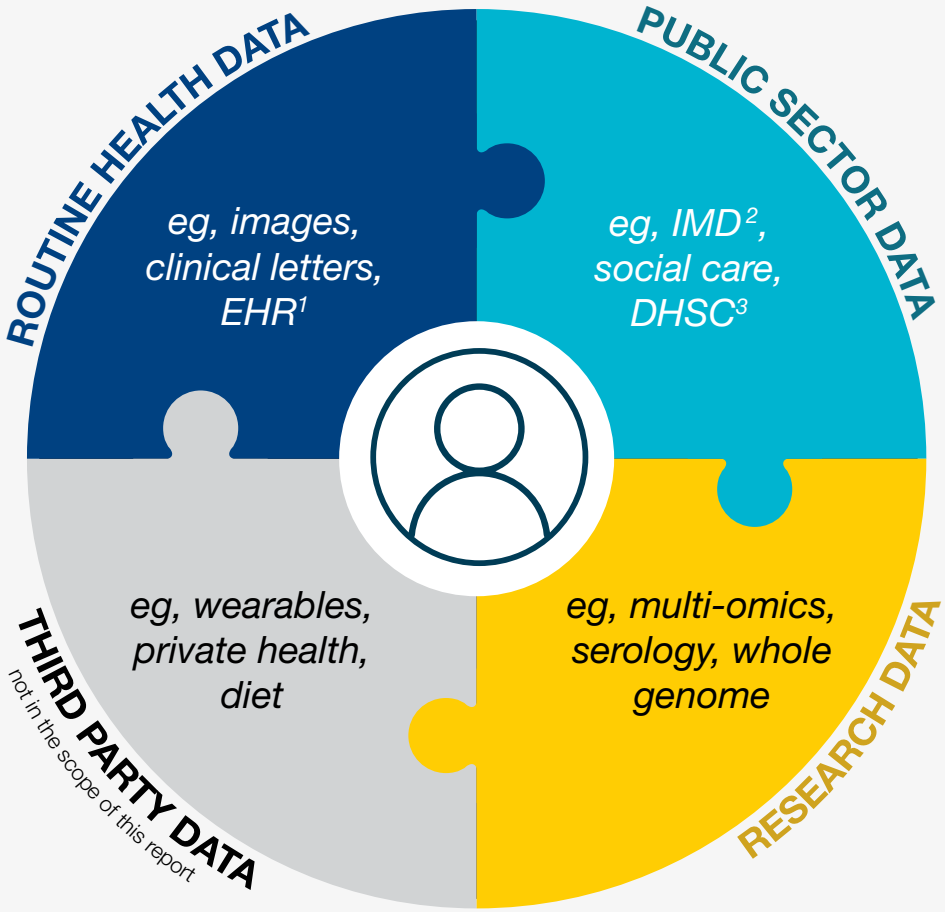
This section describes the current state of the UK health data ecosystem, based on a literature review, submitted evidence, stakeholder interviews, and workshops. It begins with the data itself: data types and values, the nature of fragmentation, and how context differs across the four nations. It then examines how data flows from source systems to researchers, the technology components at each layer, and the national and sub-national assets built. This landscape description provides the foundation for the gap analysis that follows.

Characteristics of health data

Data about individuals

This report focuses on three categories of sensitive individual-level data (Figure 1): data collected during routine healthcare; non-healthcare data about an individual (e.g., social care, deprivation, employment); and data collected through consented research. Data collected by individuals about their own health (for example, through wearables and health apps) is considered out of scope.

Figure 1. Domains of health data about an individual, with example modalities



1. Electronic Health Records, 2. Indices of Multiple Deprivation, 3. Department of Health and Social Care

Routinely collected health data includes all information stored to deliver patient care within the NHS and HSC, and forms the patient record. This includes prescriptions, diagnostic results, imaging, structured electronic health record (EHR) entries, and free text in clinical notes and letters.

Non-health public sector data aligns with an HDRS core capability in linking health data with health-relevant data for greater research impact. Health outcomes are shaped by factors beyond healthcare: housing, education, employment, income, social care, and environmental exposures. Public sector data on these determinants are held by local authorities, government departments, and agencies.

Non-routinely collected research data require patient consent. This data are collected either for observational studies (longitudinal cohorts) or interventional studies (clinical trials). Such studies may collect data specifically for research purposes – tissue samples, patient-reported outcomes, additional investigations – and may link to participants' routine care records where consent permits.

What determines data value for users?

Data value is relative to purpose. A clinical trial sponsor, an epidemiologist, an AI developer, and a regulator each use data differently. However, certain characteristics consistently determine value across use cases. For the HDRS, these define what the service must deliver to attract users. Characteristics can be understood through five dimensions:

- **Volume:** The scale of data available. A study requiring 10,000 patients with a rare cancer subtype needs population-scale coverage with sufficient follow-up to capture sufficient cases. An AI model trained on genomic data may need petabytes of data to identify an appropriate signal. The volume of data in healthcare is measured in bytes, people, and longitudinality.
- **Variety:** The richness of information captured. A clinical trial identifying eligible patients needs staging, biomarkers, and prior treatment lines. Outcomes research requires event occurrences and clinical histories. Variety means combining sources and modalities that may be siloed: linking genomics to clinical records, imaging to pathology, and prescribing to outcomes. Shallow data limits what questions can be answered.
- **Velocity:** The timeliness of data. Post-market surveillance detecting a safety signal requires data flows measured in days to weeks. A pharmaceutical company deciding whether to locate a trial in the UK needs an up-to-date view of the population. In the UK landscape, velocity varies dramatically, with some linked assets being more than a year out of date.
- **Veracity:** The trustworthiness of data. This is confidence that the data accurately reflects the capture of actual clinical events and that the variables in the dataset are complete for all patients. Linkage of parts of a patient's longitudinal record is essential for fully representing a patient's journey. This 'completeness' is a key aspect of data quality. UK data suffers from high missingness. For example, less than 7% of national outpatient data has a diagnosis code.
- **Value:** Is subsequently the result of these dimensions interacting with a use case. Data that is perceived as low value to one research group may be high value to another. However, in general, high volume, high variety, high velocity and high veracity data will be more valuable as it can be used for more use cases.

Two additional characteristics are essential when considering the value of data for a potential user.

- **Provenance:** For research to inform regulatory decisions (drug approvals, device authorisation, label extensions), data must have an auditable lineage. Regulators, including the Medicines and Healthcare products Regulatory Agency (MHRA), Food and Drug Administration (FDA), and European Medicines Agency (EMA), require the ability to trace transformations from the source system to the analytical dataset. This is non-negotiable for real-world evidence submissions. Some UK assets meet this standard. The CPRD, for example, maintains documented transformation logic and enables validation against contributing GP practices. Researchers can understand exactly how source data became the dataset they analyse. Many other UK data assets do not. For the HDRS, provenance is not a technical nicety but a commercial requirement.
- **Representativeness:** Research findings are valuable when they generalise. A treatment effect demonstrated in a study population should predict the effect in the broader population that will receive that treatment. Data that systematically excludes or under-represents population groups undermines this. All datasets face this challenge, but it is particularly prevalent in consented research cohorts. Recruitment historically over-represents white British populations and less deprived areas. UK Biobank's participants, for instance, are not demographically or ethnically representative of the general population[164]. UK data should be valued for its diversity, but only if better population representation is prioritised in HDRS development[16].

Lack of interoperability and fragmentation at the source

The UK's routinely collected health data does not exist in a unified system awaiting connection[11,35]. It is distributed across thousands of independent organisations, stored in hundreds of different software systems, structured in incompatible formats, and governed by different legal frameworks across four nations. This fragmentation at source determines what data is available for research, how it can be accessed, and what infrastructure is required to use it. The HDRS cannot redesign this landscape; it must work within it – building on existing centres of excellence and best practice.

Table 2. Data systems landscape for the four nations

Modality	England	Wales	Scotland	Northern Ireland
Primary care EHR	EMIS, TPP	EMIS	Vision, EMIS	EMIS
Secondary care EHR	>32 providers	Welsh clinical portal	TrakCare	Encompass (Epic)
Lab results	Large numbers of bespoke systems	LIMS Cymru (Citadel)	Citadel, Magentus	Core LIMS (Clinisys WinPath Enterprise)
Genomics	Multiple providers			
Images	Multiple PACS systems	National Archive (RISP)	National PACS/ Scottish Medical Imaging service	Enterprise VNA (NIPACS+)
Letters	Multiple providers	Docman/ Welsh Care Record Service	Docman/SCI store	EpicCare Link

Where individual-level data originates

Diverse settings each have distinct systems and data characteristics.

Primary care is the most consolidated source. Three vendors (EMIS, TPP, Vision) cover almost all GP practices across the four nations and have established extraction pathways. These systems maintain longitudinal records that often span decades.

Hospital care is the most fragmented. Hospitals operate EHRs, laboratory information systems, pharmacy systems, radiology archives, and speciality databases. Over 30 EHR vendors operate across hospital trusts, each storing data in proprietary formats. Data also flows to national collections (such as Hospital Episode Statistics) through professional clinical coding.

Speciality registries and audits capture detailed clinical information for specific conditions, often through manual clinician entry, providing depth unavailable in routine extracts.

Biobanks and research cohorts collect data prospectively from consented participants. UK Biobank (500,000 participants), Genomics England (180,000+), and Our Future Health (targeting 5 million) combine questionnaires, physical measurements, biological samples, and genomic sequencing.

Non-health public-sector data on housing, education, employment, income, social care, and environmental exposures are held by local authorities, government departments, and agencies, using diverse systems with inconsistent standards.

Differences across the four nations

Healthcare is devolved, resulting in variation across the four nations in organisational structure, legal frameworks, and technical infrastructure. Table 2 summarises source systems. Key differences are outlined below.

England (56 million population) has the most heterogeneous landscape. Over 200 hospital trusts and 6,000 GP practices make independent procurement decisions. Interoperability relies on messaging standards rather than shared platforms. England operates a unique National Data Opt-Out for research and planning purposes, under which NHS numbers can be checked against a service.

Scotland (5.5 million population) manages patient care through 14 health boards, 200+ hospitals and 900+ GP practices. In 2001, the Scottish Care Information (SCI) programme was set up to develop online clinical information stores. This included a store (SCI store) for clinical documents, a specialist diabetes collaboration and gateway and a format for transferring data (SCI XML). This, combined with a universal Community Health Index (CHI) number (equivalent to an NHS number in England, but also used for social care) has allowed Scotland to accumulate nearly 25 years' worth of referrals, discharge letters, laboratory results and GP summaries, with convergence towards a single provider of secondary care EHR software (TrakCare) and a single GP provider (Vision).

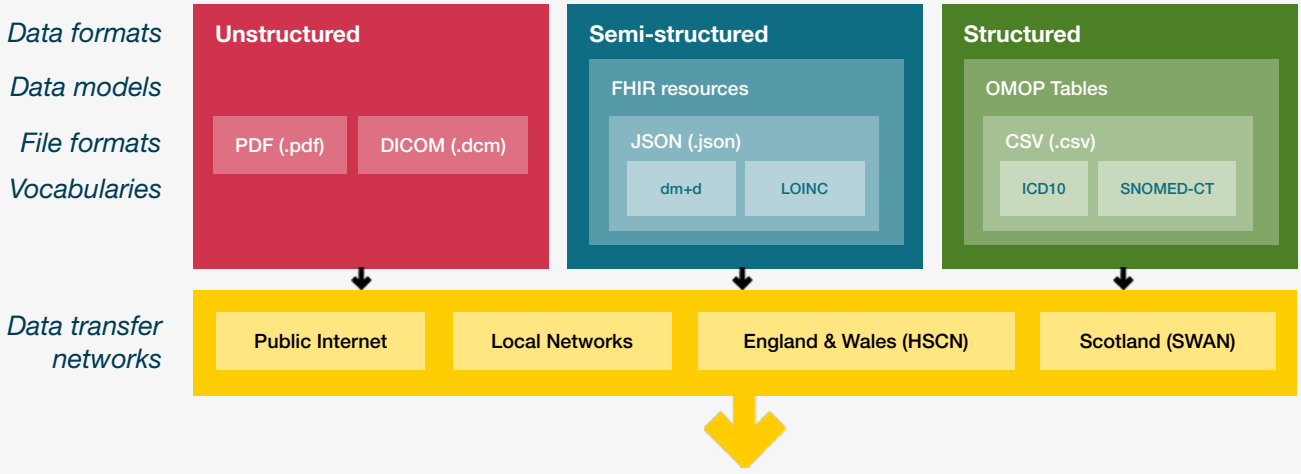
Wales (3.1 million population) has a centralised approach to health data management across its seven health boards. Unlike England's fragmented market, Digital Health and Care Wales (DHCW) acts as a central digital service provider, procuring and managing national systems across all seven health boards, achieving interoperability through shared platforms.

Northern Ireland (1.9 million population) operates a Health and Social Care (HSC) system rather than a national health service, comprising five hospital trusts and a single ambulance trust. Historically, digital maturity has progressed more slowly than in the other UK nations; however, recent transformation has been significant. The Encompass programme has deployed Epic as a single regional EHR, alongside national imaging and digital pathology initiatives, creating a unified direct-care digital architecture. There is an absence of a statutory secondary-use framework in Northern Ireland, which constrains the use of unconsented data for research and limits routine linkage between primary and secondary care.

Why interoperability standards have not solved this

For data to move between systems and be combined for research, it must conform to formats that receiving systems

Figure 2. File formats, data models, vocabularies and networks combine to deliver interoperability



can interpret. Interoperability standards attempt to establish these conventions at four levels:

File formats specify how data is encoded and stored, and define how systems can read each other's files. Structured data is often found in tabular formats. Digital Imaging and Communications in Medicine (DICOM) provides the universal standard for medical imaging[36]. FASTQ and BAM/CRAM store genomic sequencing data[37]. PDF and Microsoft Word remain the default for unstructured clinical documents.

Storage and exchange mechanisms define how data is held and how it moves. Clinical data sits in proprietary database systems (e.g., Epic's Clarity/Caboodle, Cerner's backend, TPP's SystmOne). Data may move via exports (CSV, JSON, XML), messaging protocols (HL7v2), or APIs (FHIR). Open table formats (Delta Lake, Apache Iceberg) represent an emerging alternative: vendor-neutral formats that different analytical engines can query directly. All these exchange mechanisms must operate across a network. Local networks at individual sites, private networks used across health systems and the public internet are all relied upon to transfer information.

Data models specify how information is structured within those formats – what fields exist, how they relate, and what each represents. FHIR, as a messaging protocol, also defines data structures for information exchange[38]. OpenEHR provides an archetype-based approach to clinical data modelling[39]. The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) is an example of a standard framework for observational research[40]. The NHS Federated Data Platform uses its own common data model[8].

Vocabularies specify a language for clinical concepts within these structures. SNOMED CT provides comprehensive clinical terminology[41]. ICD-10 classifies diagnoses and causes of death[42]. OPCS-4 codes surgical procedures[43]. The NHS Dictionary of Medicines and Devices (dm+d) standardises medication and device references[44]. Shared vocabularies are vital for ensuring data is interpreted uniformly.

A problem is not the absence of standards, but the lack of consensus. At least nine global initiatives define healthcare data standards[45], sometimes in competition for the same purpose. No single standard can cover all data types and purposes[46].

Lack of adoption remains the deepest problem. This is required to overcome the primary engineering challenge in health data research: source data is found in a variety of structures and formats that are proprietary to different software vendors (e.g., Epic, Cerner, EMIS, TPP) and are also dependent on local infrastructure and organisation-specific configurations.

Implications for the HDRS

Across all four nations, there is no standardised approach to extracting and combining data from source systems for research. Each system and data type requires bespoke technical arrangements. It is currently impossible to systematically identify which sources hold records for a given individual across primary care, secondary care, speciality systems, registries, and research datasets – within or across the four nations of the UK.

The HDRS cannot solve this issue, but must build mechanisms that work across this heterogeneity. Where HDRS pursues standardisation, it should consider focusing on standards that address critical gaps in research capability, demonstrably improve service delivery, and have a lower barrier to widespread adoption.

How health data flows to users

In the UK, health data for research does not currently flow through a single pathway. Depending on the data type, geographic scope, legal basis, and intended use case, data may pass through different combinations of processing steps and access mechanisms. Five layers describe how data may move from generation to end-user analysis and one layer that determines how end-users access the data.

Source systems layer

Source systems are the clinical and administrative systems where health data is first generated. As discussed, this layer is highly heterogeneous. Across the NHS, there is no single dominant system, and even within a single trust, multiple systems typically coexist[47]. This fragmentation at source means that no universal extraction approach exists and that complexity propagates through all downstream layers[48].

Source data processing layer

This layer extracts and transforms data from source systems. Raw data can be extraordinarily complex (for example, tens of thousands of tables are in the Epic EHR database). This stage makes data more accessible and amenable to analysis, but may fall short of curating a dataset that is easy to use for research.

Data engineering uses programming code to extract, clean, validate, and transform data. Access to source databases depends on vendor consent, local technical capability, and expert understanding of database structures. Human-mediated processes add value but also constraints – clinical coders can translate documentation into standardised codes, but throughput limitations affect timeliness, and this data remains sparse. Emerging approaches use optical character recognition and large language models (LLMs) to extract structured data from unstructured documents at a quality comparable to human coders[49].

There are, however, limitations to the use of such methods. For real-world evidence and regulatory submissions, transformation logic must be documented and auditable from source to analytical dataset for the purposes of establishing provenance. The use of blackbox AI in this process may undermine the ability of organisations to meet these regulatory transparency requirements.

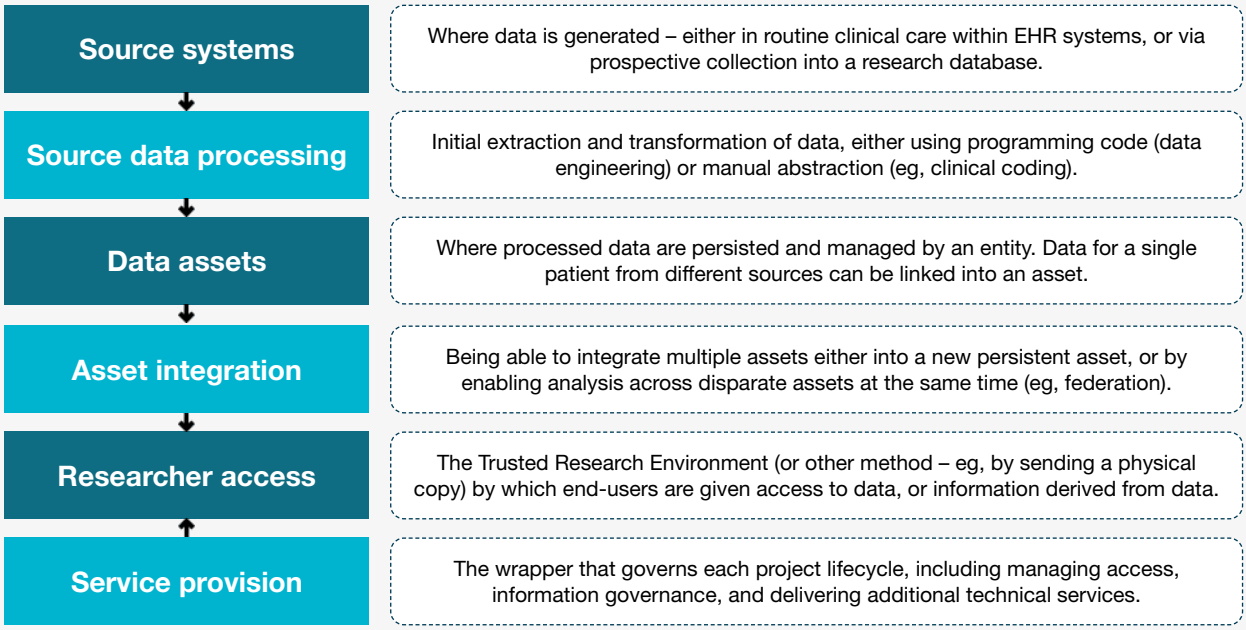
Data asset layer

The data asset layer is where processed data is persisted in a managed format that supports researcher access.

Linkage is a critical function at this layer: combining data from different sources for the same individual. This typically relies on a persistent identifier, though probabilistic matching may be required[50]. Linkage may occur at the point of ingestion (data arriving pre-linked) or within the asset itself (linking datasets after ingestion).

Data assets vary in volume and variety. Some hold single data types, others integrate multiple sources into linked longitudinal views for large populations. Not all assets are housed within secure research environments: many remain

Figure 3. The six layers of research data infrastructure



Layers of research data infrastructure. The research data pipeline flows from source systems to researcher access via secure environments. The service provision layer supports researcher access to research data.

on hospital servers, others in less secure settings, or distributed via data releases to personal devices[4].

The storage size of data assets varies dramatically by modality, with implications for infrastructure design. England's structured primary care data for 57 million people is approximately 3.6 terabytes (0.06MB per person)[4]. UK Biobank holds 30 petabytes for 500,000 participants (60,000MB per person)[51] – a millionfold difference. These scale differences affect how data can be processed and integrated with other assets.

Asset integration layer

This layer provides mechanisms to join separate data assets, enabling analysis across larger populations. It is not universally present – most UK assets operate in isolation. Discovery infrastructure is part of this layer, enabling researchers to identify relevant data assets for a study.

Where asset integration exists, it takes several forms: (1) replication into a new environment where data persists as a new asset; (2) temporary transfers into an environment for a specific analysis; or (3) federated analytics by sending code to distributed data assets.

All three approaches can be observed in the UK – for example, in England's sub-national Secure Data Environments (SDEs) and Data Safe Havens in Scotland, where research cohorts found in different sites are brought together, or OpenSAFELY, which federates queries to separate data assets held by two GP EHR vendors (EMIS and TPP).

Integration requires addressing the interoperability challenges described earlier. A standard approach for discovering and integrating assets does not yet exist. Standard data models and vocabularies are prerequisites for federated analytics. When assets are brought together, further engineering may also be necessary to standardise data models to enable consistent analytics.

Research access layer

This layer describes environments and interfaces through which approved users access and analyse data. This is where data flows terminate, and research value is generated[52].

TREs represent the accepted model for secure access: controlled workspaces where users analyse data without extraction, with restrictions on data export and audit trails of activity[5]. TRE maturity varies considerably. Some offer high-performance computing for machine learning, while others provide only basic desktop functionality with limited software customisation[53]. For AI workloads, compute capability is a particular constraint[54]. Training models on imaging or genomic data requires GPU clusters' storage capacity and data throughput that most current TREs cannot provide.

Despite policy direction toward TREs, anonymised data release to end-users remains common[53]. This offers flexibility but provides weaker security and limited auditability, and risks relinquishing the value of anonymised data assets to third parties.

In federated analysis, users submit code for execution across distributed assets and receive only aggregated results[55]. This requires prior understanding of data models at each node, as direct 'eyes on' data inspection is not possible.

This layer must also support research outputs through disclosure control processes, ensuring released results do not risk re-identification as well as logging and audits of processing activity.

Service provision layer

This layer encompasses the processes enabling a research project to progress from conception through data access to dissemination. While preceding layers describe technical data flows, this layer describes how researchers and data providers interact.

Access pathways vary considerably. At one extreme, access may be obtained informally through existing relationships or established research groups[5]. At the other end, digital platforms manage requests with minimal human interaction. Most pathways fall between, combining limited automated systems with substantial manual coordination.

Common steps are typically required regardless of the pathway. Researchers must confirm the suitability of the data for their question. Data providers must assess alignment with approval policies. Ethical approval must be obtained. Commercial agreements must establish value sharing. Practical arrangements must cover data transfer or access, analysis, conduct, and the disclosure of results.

For the HDRS, this layer determines user experience. The user stories demand predictable timelines: rapid access for pharmaceutical customers, rapid feasibility for trial sponsors. Current service provision is characterised by variable processes, uncertain timelines, and bespoke negotiations.

Summary

This six-layer framework describes how health data moves from generation to analysis, but there are caveats.

First, data flows are not linear. Some research relies on direct transfers from source systems to laptops, bypassing intermediate layers. Many assets operate in isolation, lacking integration capabilities.

Second, the framework describes the current state, not a prescription for the future. The HDRS will need to interact with, improve, or potentially supersede capabilities at multiple layers.

Third, different use cases involve different combinations of layers. A clinical trial feasibility query may require only aggregated counts from multiple assets. A precision medicine study may require linked multimodal data flowing through all six layers into a high-performance computing (HPC) environment. The HDRS will need to accommodate this diversity.

Technology components

The data flow framework identifies what must happen at each layer. This section examines the technology components that enable these functions, their maturity in the UK, and implications for the HDRS.

Source data connections (source systems layer)

Source data connections are the technical mechanisms used to extract data from clinical and administrative systems. These include direct database connections, application programming interfaces (APIs), messaging standards (such as HL7 and FHIR)[56], and file-based extracts.

GP system vendors (TPP, EMIS, Vision) have established extraction pathways used by multiple downstream assets, though access terms and pricing vary. Hospital EHR systems present greater heterogeneity. Modern systems may expose FHIR-compatible APIs, but these are designed for transactional messaging rather than bulk research extracts, and are not compatible with all data in system backends. Most hospital data extraction requires direct database access or manual abstraction by clinical coders.

No standardised extraction approach exists across the NHS. Capabilities depend on local technical expertise and vendor relationships. Third-party suppliers have developed integration patterns that are proprietary intellectual property, leading to duplicate procurement across sites. For HDRS, source connections impose a binding constraint on the data that can flow into the system.

Data pipelines (source data processing layer / data asset layer)

Data pipelines automate the extraction, cleaning, validation, and transformation of source data into research-usable forms. Different types of research may require different degrees of transformation, occurring at earlier or later stages of the pipeline.

Pipeline maturity varies substantially. GP data pipelines extract data from well-established APIs. Some assets (such as CPRD, SAIL) operate mature pipelines developed over years, with documented transformation logic and quality assurance. Regional SDEs are at earlier stages, with few having automated ingestion from constituent trusts. Most rely on ad-hoc transfers. For most NHS trusts, pipelines primarily serve reporting, not research. Most valuable data remains in 'raw' form.

A common constraint is expertise. Building robust pipelines requires data engineering skills scarce across the NHS and academic sectors. Where pipelines exist, they often depend on key individuals, creating sustainability risks. SeRP's dedicated technical team of 60 represents an exception; most assets operate with significantly smaller capacity. For the HDRS, pipeline capability directly determines data value.

Data linkages (data asset layer)

Data linkage combines records from different sources for the same individual, using deterministic matching on NHS/CHI/HSC number or probabilistic methods where identifiers are absent[50].

Linkage capability is well established across several national assets. NHS England operates the Master Person Service for national dataset linkage[57]. SAIL has developed sophisticated probabilistic methods for non-NHS sources[58]. Scotland operates a dedicated CHI linkage team[59].

Constraints are primarily governance rather than technology. Linkage typically requires approval from the NHS Confidentiality Advisory Group (no equivalent exists in Northern Ireland) or explicit consent. For assets seeking to expand linkage to new sources, governance represents the binding constraint. Reliable linkage underpins the comprehensive longitudinal records that user stories demand.

Data storage architectures (data asset layer)

Data warehouses and data lakes provide the storage layer for data assets. Warehouses impose structured schemas optimised for query performance; lakes store data in raw or semi-structured forms, offering flexibility at the cost of requiring transformation at query time.

Most established UK assets use relational databases (e.g., SQL Server, PostgreSQL, MySQL), reflecting their origins in structured administrative data. Cloud-based architectures (Databricks, Snowflake) are increasingly adopted by newer assets, offering scalability for diverse data types. Where requirements are predictable and volumes are high, hybrid infrastructure that include on premise compute and storage (as at Wellcome Sanger, Francis Crick) can be more cost-efficient.

For the HDRS, architectural choices affect the feasibility of integration. Assets using common technologies integrate more readily than those with proprietary storage.

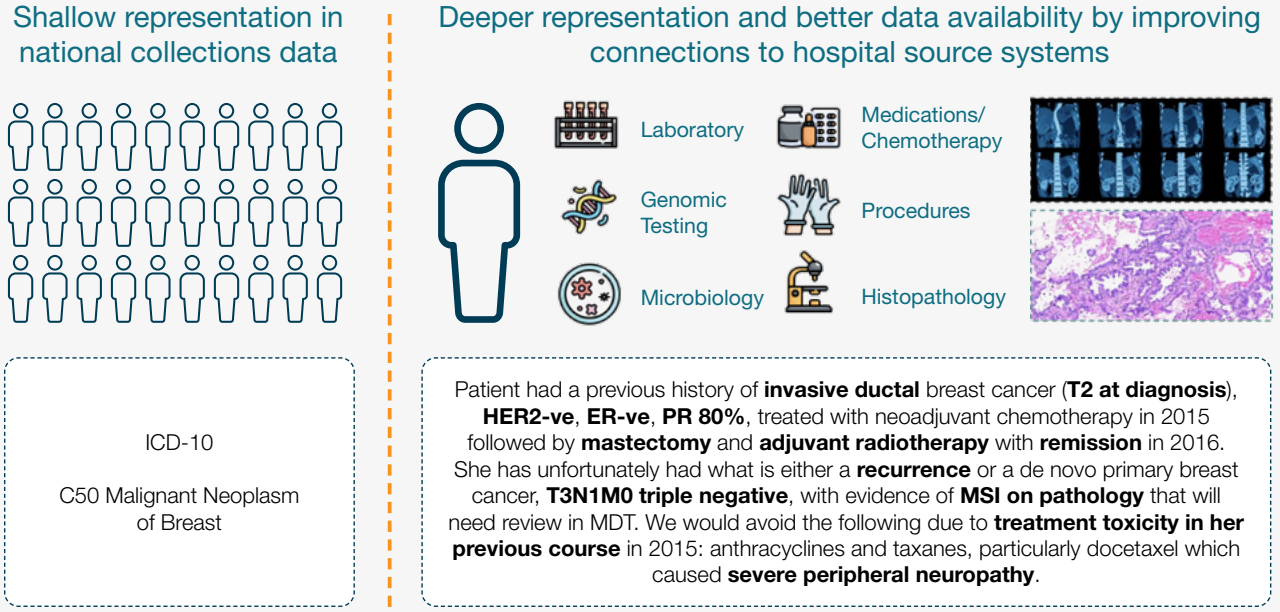
Unstructured data and natural language processing (source data processing layer / data asset layer)

Unstructured data (clinical notes, letters, reports) contains valuable clinical detail absent from coded fields. Natural language processing (NLP) using LLMs can extract structured information at scale[15].

UK capability remains relatively immature. Academic groups have developed disease-specific extraction tools. Some trust-level assets have implemented NLP pipelines. Genomics England employs NLP for data enrichment. Commercial technologies include John Snow Labs[60], CogStack Ltd[61], and Akrivia Health[62]. An NHS R&D project is funding the deployment of multi-site LLMs for cancer data enrichment[63]. Primary care unstructured data remains largely inaccessible, locked behind per-project governance agreements with GP data controllers and stored in vendor environments that require substantial investment to unlock.

For the HDRS, unstructured data represents a substantial untapped resource. Scalable processing capability is essential to support user stories that require clinical depth beyond coded fields.

Figure 4. Illustration of the context provided by including unstructured data



Although patients may seem similar in a national collection when identified by an ICD-10 code, local information captures greater depth, representing patients with far more detail. With proper integrations, this data can be surfaced to researchers – including prescribing and pathology data, multi-modal data, and concepts from free text.

Common data models (data asset layer)

Common data models standardise data representations, enabling consistent analytics across assets with different native structures.

OMOP, maintained by the OHDSI collaborative, is an example of a widely adopted international standard[64,65]. The OMOP network is more mature in Europe and the EMA is making use of the standard alongside 180 data partners to accelerate regulatory grade research on the continent. UK OMOP adoption is growing with pipelines developed across the SDE Network in England, Data Safe Havens in Scotland, CPRD and UK Biobank. However, OMOP is not a silver-bullet solution, and has limitations. Vocabularies are US-centric and do not employ UK standards (e.g., dm+d for medications) as defaults. The restricted data model may abstract away information from complex medical concepts. The format is designed for 'eyes-off' analysis, not data exploration.

Any transformation to a common data model requires substantial and ongoing effort: mapping local codes to standard vocabularies, resolving semantic ambiguities, and validating outputs. Few UK assets have completed comprehensive implementations as part of automated pipelines. Where UK sites are part of international networks, these are project-specific academic networks to coordinate separate analysis at each site, and are not set up for broad end-user access as a service.

Despite these limitations, for federated analytics approaches, common data models are essential: without a consistent structure and semantics across nodes, federated queries cannot operate, or aggregate results cannot be interpreted reliably[66]. An example of this is the Cohort Discovery Tool (HDR UK) which leverages a minimal OMOP data model to federate feasibility queries across multiple data assets without moving the data[67].

Data discovery (data asset layer / asset integration layer)

Data discovery enables researchers to identify what data exists and assess its suitability for their research question[30].

The HDR UK Innovation Gateway provides the only UK-wide discovery infrastructure[68], cataloguing dataset descriptions, structural metadata lists, and data access request information. It currently contains over 1,000 datasets from 103 data controllers.

Larger data providers also provide support for data discovery with detailed documentation or the publication of synthetic data (e.g., CPRD, Research Data Scotland, OpenSAFELY) enabling deeper feasibility assessment.

Feedback from workshops and interviews indicated that current discovery mechanisms were insufficient to determine whether data was fit for purpose[11,35,69]. There is also a question as to the utility of automated solutions to

data discovery given the relatively large proportion of research requests that begin with broad questions rather than detailed protocols.

Federated analytics / learning infrastructure (asset integration layer)

Federated analytics executes queries across distributed data assets, returning aggregated results without centralising raw data[55]. This requires: data held to common specifications at each node, software and compute capability at each node, orchestration infrastructure for secure information transfer, and a hub for user interaction.

OpenSAFELY operates as a production 'eyes-off' primary care platform[70]. Beyond this, UK federated infrastructure remains largely at the demonstrator stage with open source approaches such as DataSHIELD[71] trialled in the SDE network. Reported barriers to success include a lack of transformed data at nodes, variable compute capability, and general technological immaturity in available platforms.

Federated learning is a type of federation where locally trained model weights are aggregated centralling to train new machine learning models. This requires a lower degree of standardisation at each node, but requires substantial compute capability, as smaller models are first trained locally.

Fabric infrastructure (asset integration layer)

Data fabrics provide orchestration across distributed assets, enabling discovery, access, and analysis through unified interfaces while keeping data in its native locations. Unlike federated analytics (which executes specific queries across nodes), a fabric layer coordinates the full research workflow: finding data, requesting access, and conducting analysis, regardless of where data physically resides [76,77].

Although relatively common in industry settings, no comprehensive data fabric currently operates across UK health data assets. However, increasing numbers of data assets, including English SDEs, NHS trusts, and SAIL/SeRP are hosted in fabric-compatible cloud technologies.

TREs (research access layer)

TREs provide secure workspaces where approved researchers can access and analyse data without extracting it[78]. They typically offer virtual desktops, analytics software, access controls, audit logging, and data disclosure controls. As discussed in 'research access layer', TRE maturity varies considerably.

A standard architecture (SATRE) has been proposed as a blueprint, with two TREs (The Alan Turing Institute's Data Safe Haven and the University of Dundee's TREEHOOSE) evaluating against it[79]. A DARE UK project is developing a deployable SATRE-compliant TRE (K8TRE), but this remains in development. TRE infrastructure is often outsourced from a provider, such as Lifebit[72], Aridhia[159], BC Platforms[160] and DNAnexus[80], or through a hyperscaler offering (such as Azure or AWS). University-based teams, either via BRCs or ARCs, or the data platform team at SeRP, are examples of self-hosted TREs.

TREs typically sit in separate organisations from data controllers and data assets. As richer data flows through automated pipelines, proximity to data and access speed become important considerations.

Compute resource (research access layer)

Compute infrastructure encompasses processing resources within research environments, from basic desktops to high-performance computing and GPU clusters.

Capability varies substantially across TREs. Some offer only basic virtual desktops with limited memory. Others, particularly at academic institutions, provide high-performance computing access, including GPU clusters. Cloud environments offer scalability at a higher cost.

UK supercomputers (Dawn[81], Edinburgh Parallel Computing Centre[82], Isambard-AI[83]) offer potential for large-scale analysis, but present constraints for secure data usage including the inability to persist data or manage analytical project lifecycles. The DARE UK FRIDGE project aims to address this through a TRE 'extension' mechanism, but has not reached production[84].

We have found no evidence that any significant NHS data studies have run on the UK supercomputer network.

Application workflow (service provision)

The process through which a potential research project is taken from request through to data provision is duplicated across the ecosystem. Historically it has been a back and forth of emails and forms but there are examples where technology has supported the standardisation of this service. These solutions help the applicant understand what stage they are in the process, how many more stages there are left, the actions required, and timelines expected to get to completion. This can include the process for aligning health research authority application, ethics, consent, funding, data access forms and commercial agreements. For clinical trials this can include the planning and accrual stages. This process is not always linear and no technology supports the entire process. This is serviced by a range of software applications that include Salesforce, Zoho, Jira and YouTrak.

Cohort feasibility (service provision)

Determining whether a research project can proceed requires investigation of the data. In some cases this can only be done through experts querying and linking data but there are also technology components used to automate the process. These automated approaches can be delivered through a federated approach (such as the HDR UK Cohort discovery tool) or centralised service such as the DigiTrials feasibility tool or Mauro data mapper. In all cases the tools aim to reduce the time taken to understand whether the data is available for research and the cohort is large enough to answer the research question.

Supporting components

Several supporting functions are essential for HDRS operation:

- Data provenance recording has been described as an essential feature of research data[85]. Documentation of data sources and transformations, along with sufficient metadata, is critical for regulatory submissions, where data accuracy must be auditable[86].
- Data minimisation prevents all data from a data asset being shared with a researcher. In many cases access will only be approved for a subset of the data based on the inclusion and exclusion criteria of the research. The data then must be minimised by the team so that only the data that the researcher has permission to analyse is provided.
- Authentication provides the identity management and authentication tokens required to access the research environments and then access the data that is connected to these research environments.
- Repeatable analytic pipelines accelerate research through de-duplicating analytical processes. These are shared through repositories of code, phenotype libraries or documented processes. These prevent multiple slightly different approaches for achieving the same outcome.
- Pseudonymisation approaches remove information from the data that would be considered identifiable. This includes names, dates of birth, postcode, and NHS numbers. The data is not considered anonymous but that the risk of re-identification is beneath an acceptable tolerance.
- Disclosure control or output checking ensures that outputs released from secure environments do not risk re-identification and comply with data-sharing terms[87,88]. Most processes are manual, via an 'airlock team'. Automated approaches, particularly for AI model weights, currently exist in the research space[89].
- Safe user registries maintain accredited lists of approved researchers and organisations, enabling portable credentials across data controllers. A digital registry to track these credentials is being developed by HDR UK[90].
- Archiving is required to meet regulatory data retention requirements. Guidelines recommend archiving trial data for at least 25 years[91]. Archiving snapshots of analysis datasets, and keeping code used to create datasets, are ways to ensure reproducible and auditable research outputs, and a recommendation of previous reviews[5].
- Active surveillance supports regulatory requirements for rapid reporting of serious adverse events identified during trials or real-world evidence studies, and requires queries to be automatically executable on high-velocity data[92]. This is a technological challenge across all levels of current UK infrastructure and also represents a high-value capability.

Summary

The components described in this section span the data flow framework, from source system extraction through to research access environments. Across the UK, components vary in maturity. Components closer to source systems (connections, pipelines) are typically implemented alongside individual data assets. Integration layer components (federation, fabric infrastructure) are less mature, with limited production-scale deployment.

For the HDRS, this landscape indicates where investment can build on existing capability versus where foundational development is required. A national research service cannot operate purely at the access layer. It must engage with components across the full data flow, with clear boundaries defining HDRS's responsibility at each layer.

National and sub-national data assets

The UK health data landscape comprises diverse data assets that have emerged through different funding streams, governance arrangements, and technical approaches over decades. Rather than cataloguing every asset, this section groups them by the health information they hold, the technical frameworks they use, and the end-users they serve, identifying examples that offer instructive models for HDRS development. This section is not intended to be a comprehensive catalogue.

The scope captures assets with population coverage of at least an NHS trust catchment area, with current or emerging capability to support research, and where data sits a single step from source systems. This excludes the long tail of curated cohort datasets.

Sub-national SDEs

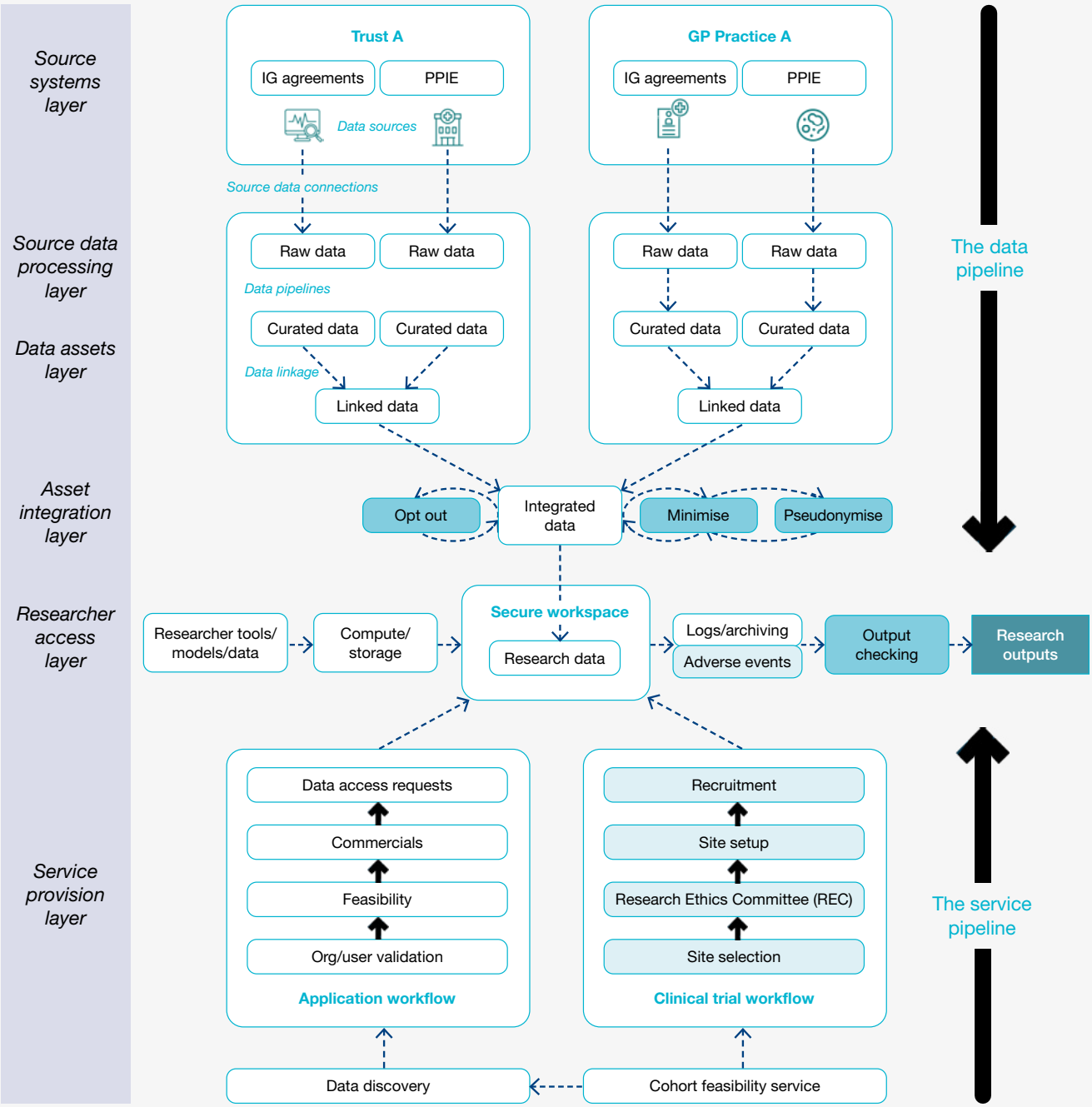
England's eleven sub-national SDEs, funded since 2022, represent a top-down intervention to create regional infrastructure that integrates data from multiple NHS organisations within a geographic footprint[9].

These environments have developed their own infrastructure, resulting in heterogeneous technology and data integration strategies across the network. While SDEs aim to federate, they currently operate as isolated islands with no standardised mechanism for cross-SDE analysis.

SDEs developed on top of existing relationships and infrastructure within an NHS trust and universities have developed deep integrations with acute systems and multimodal data offerings. This includes Thames Valley and Surrey (Oxford University Hospitals), Wessex (University Hospital Southampton), West Midlands (University Hospitals Birmingham), Yorkshire and Humber (Bradford Teaching Hospitals) and East of England (Cambridge University Hospitals). Some of these organisations are bringing in imaging and genomics projects alongside specialist acute datasets.

SDEs without this foundation have developed geographic coverage that links primary care data to commissioning datasets or to existing linked datasets within a region. This group includes: Kent, Medway and Sussex; the North West; London; and the South West. Many of these SDEs are now building on top of primary care data collections with direct access to acute trust data across an ICB.

Figure 5. The components that support the six layers of infrastructure



Key

Components	Processes	Clinical trials related

Supporting components for research data infrastructure layers. This illustrates components of the data pipeline and the service pipeline. Components are defined in the Technology Components section of this paper.

The regional model balances national coordination with local governance: each SDE operates under its own data controller arrangements while adhering to common accreditation standards. This approach has enabled data sharing agreements between SDEs and over a third of English GP practices. Eleven of twelve SDEs have obtained Confidentiality Advisory Group (CAG) approval for patient-identifiable data, enabling opt-out checking and linkage via NHS number.

The intermediary nature of SDEs creates challenges for provenance. Engineers may be unable to trace lineage to primary sources, particularly where third parties provide intermediate pipelines[93]. Service delivery maturity varies considerably, from several years of research delivery to first customer pilots.

Population-scale primary care linked databases (e.g., CPRD)

Population-scale primary care databases are among the UK's most established research assets, dating back decades. These operate at the data asset layer, receiving extracts from GP systems, performing extensive data engineering, and linking to secondary care, mortality, and other national datasets.

The CPRD is an exemplar here, providing longitudinal primary care records that can be linked to Hospital Episode Statistics (although not for clinical trial recruitment due to information governance barriers) and other datasets, covering around 30% of the UK population[11]. Since 1988 CPRD has contributed to 3,779 peer reviewed publications in the UK and internationally[161]. CPRD operates under MHRA oversight, with established data-sharing relationships to participating GPs, and a mature commercial access model. CPRD offers specific clinical trial services: 'CPRD Sprint' that facilitate patient case-finding and recruitment, acting as a bridge between data and GP practices[7]. This close GP relationship enables data quality validation against source systems and trial recruitment through practitioners, increasing success rates and avoiding per-project governance approvals. CPRD has recently implemented an Azure TRE, moving away from historical anonymised data releases.

Coverage depends on GP practice opt-in, which may introduce selection bias, though representativeness remains sufficient for generalising real-world evidence findings.

CPRD and other primary care databases typically face data richness limitations, specifically lacking secondary care prescribing, laboratory and other biomarker data[69].

OpenSAFELY code-to-data platform

OpenSAFELY represents a distinctive architectural approach that emerged during COVID-19[94], demonstrating analysis of sensitive primary care data without researchers viewing person-level records. Researchers write code using synthetic data, submit it for execution within GP vendor environments, and receive only aggregated outputs, subject to disclosure controls.

This model achieved over 98% English GP coverage by working directly with TPP and EMIS back-ends, avoiding data extraction entirely. Rather than data moving through processing layers into a separate asset, analytical capability is brought to data at the source. OpenSAFELY has developed ehrQL, a specialised query language that facilitates code reuse.

The approach can reduce friction for technically proficient researchers (through documentation, synthetic datasets, and tool integrations) whilst maintaining data controller confidence. The software is suitable for researchers who do not require direct data access. The model depends on vendor cooperation for access and infrastructure.

OpenSAFELY remains the closest capability in England for population GP data coverage. The information governance enabling this has previously relied on COVID-19 approvals but a recent ministerial direction[125] and subsequent GP engagement has seen 92% (as of February 2026) of GP practices on-boarded for non-pandemic research.

Clinical trial recruitment platforms (e.g., NW eHealth, DigiTrials)

Data-driven clinical trial infrastructure supports feasibility assessment, site identification, and patient recruitment rather than observational research.

NHS DigiTrials, operated by NHS England, enables sponsors to query aggregate patient counts against eligibility criteria across NHS datasets (Hospital Episode Statistics and primary care dispensing) and to replicate them into a DigiTrials data asset[95]. The query interface returns only aggregate counts. DigiTrials interfaces with NHS digital communications for direct patient outreach, currently requiring Section 251 approval. The service is seeking full-service directions to become a data controller and streamline governance. The absence of complete primary care diagnostic data or secondary prescribing limits query granularity, though expansion to Mental Health and National Disease Registration Service data is planned. While in pilot, DigiTrials has supported population-scale trials "into the double digits". Unlike CPRD, DigiTrials does not rely on clinicians caring for patients to facilitate recruitment and has a comparatively low conversion rate from outreach.

North-West EHealth (NWEH) offers a commercial alternative addressing recruitment governance[96]. Its Farsite platform splits data into identifiable (for GPs) and pseudonymised (for researchers) streams[97]. Researchers query the pseudonymised view, then the system re-identifies participants after GP review. ConneXon – a community-powered organisation (founded in 2022, formerly Cambridge-based) that acts as an "ecosystem activator" for life sciences, biopharma, and health tech industries – complements this with eConsent and ePROMs, integrating direct data capture from routine care[98].

Currently, no clinical-trial feasibility or recruitment platform offers multi-region integration with deeper data for hospital-run clinical trials.

Wales SAIL and SeRP

The SAIL Databank represents Wales' national infrastructure and one of the UK's most mature integrated health data systems[78]. SAIL spans all data flow layers: receiving data from Welsh NHS sources, performing data engineering and probabilistic linkage for non-NHS data, and providing researcher access via the SeRP. SAIL holds linked primary care, secondary care, administrative, social care, education, and environmental data, though data from NHS trust source systems are not yet integrated. A service offering supports feasibility, data understanding, and cohort creation.

SeRP provides the technical infrastructure that can be deployed across multiple TREs, with a full-service offering that includes management of physical infrastructure, pipeline engineering, and TRE management. The infrastructure provides both physical and logical separation of data for different user groups. Most SeRP infrastructure is on-premise, in self-managed data centres located in Swansea. SeRP also manages cloud infrastructure on behalf of other data controllers (e.g., SeRP Canada). The technical team at SeRP numbers 60 individuals out of approximately 200. SeRP is evolving from a traditional on-premise relational database (IBM Db2) to a 'data fabric' architecture.

Scottish infrastructure (Safe Havens network, linkage services)

Scotland's foundation, described previously, supports a tiered network comprising a National Safe Haven and four regional Safe Havens. The National Safe Haven, operated by Public Health Scotland via eDRIS, covers the entire 5.5 million population and provides access to national datasets (Scottish Morbidity Records)[99] as well as integrated data on prescriptions across primary and secondary care (SCoMeD[162]). Regional Safe Havens (Glasgow, DataLoch[100], HIC Tayside[101], Grampian Data Safe Haven[102]) hold deeper, more granular, clinical data.

Access to primary care data remains operationally complex. While the technical capability to extract data exists, predominantly using Albasoft technology to interface with Vision and EMIS systems, GP practice approvals act as a constraint. DataLoch has successfully mitigated this by incentivising GP data sharing through a 'value-back' model, providing analytical dashboards to practices, but this approach has not yet scaled nationally.

Scotland has imaging and pathology data available nationally for research. The Scottish imaging archive contains 57.3 million radiology studies linked to their medical records[103] whilst the Brain Health Data initiative holds 417,000 magnetic resonance imaging (MRI) and 846,000 CT scans linked to EHR data and free text[104]. This, combined with the national TRE co-located at the EPCC provides access to high performance compute for large multimodal datasets[82].

Despite the network structure, automated federation is absent. Previous attempts (e.g., Connect 4) were largely unsuccessful. Cross-regional projects rely on manual triage and shared salt key. A salt key is a random, unique set of characters added to the end of a password to prevent hackers using precomputed tables to reverse-engineer passwords – mechanisms to pseudonymise and transfer data to a preferred TRE.

Northern Ireland infrastructure

Northern Ireland provides access to health data for research primarily through the Honest Broker Service, operated by the Business Services Organisation (BSO)[105]. Historically, the Honest Broker Service has sat above a regional data warehouse containing longitudinal records spanning hospital activity, pathology, and community prescribing. Primary care data is managed separately through the General Practitioner Intelligence Platform, which integrates data from the region's single GP system supplier (EMIS) under agreements established with all GP practices across Northern Ireland. Data is made available to researchers through the Northern Ireland TRE deployed within an instance of the SeRP.

Analytical capability is delivered through the Northern Ireland Health Analytics Platform (NIHAP)[106], a cloud-based environment hosted on Microsoft Azure and established during the COVID-19 pandemic to support population-scale analysis.

This landscape is now in transition following the rollout of Encompass, which has deployed Epic as a single EHR across all five health and social care trusts. While this represents a major advance for direct care, it has disrupted established research data pipelines and introduced additional architectural and service-layer complexity for extracting research data from Epic systems. These technical challenges are compounded by the absence of a statutory framework in Northern Ireland permitting the use of identifiable data for secondary purposes without consent, which limits linkage between primary and secondary care data and constrains the development of centralised data services.

In particular, Northern Ireland currently lacks a region-wide service wrapper equivalent to England's integrated feasibility, recruitment, and follow-up infrastructure for clinical trials. Although the underlying data assets exist, the absence of a consented data extraction and validation service places responsibility on individual research teams, creating friction for both academic and commercial studies. Stakeholders consistently identified this gap as a primary barrier to participation in large-scale trials and to forming sustained partnerships with life sciences companies.

Biobank/longitudinal cohorts (e.g., UK Biobank, Genomics England, Our Future Health, UK Longitudinal Linkage Collaboration (UK LLC), NIHR BioResource)

Longitudinal cohort studies and biobanks represent a distinct asset type characterised by deep, multimodal data on consented participants, typically including genomic, imaging, and lifestyle data alongside linked health records.

UK Biobank[12], Genomics England[13], Our Future Health[14], UK LLC[107] and NIHR BioResource[108] are just some examples.

These assets operate differently within the data flow framework: participants are recruited prospectively, and any biological samples are processed through dedicated laboratory pipelines, while NHS linkage provides longitudinal outcomes. Data engineering can encompass genomic sequencing, imaging processing, and integration of research-collected data with routine records. Each operates its own TRE.

These assets have unique value for precision medicine and epidemiological research, providing the multimodal data needed to identify disease subtypes, validate biomarkers, and support targeted therapeutic development.

UK Biobank's research access model, processing thousands of requests per year and providing data to approved researchers globally – has generated substantial research output and demonstrated commercial viability through transparent and simple charging mechanisms.

Genomics England, emerging from the 100,000 Genomes Project, provides a model for integrating genomic data with NHS care pathways and returning findings to clinical practice.

Our Future Health aims to recruit 5 million participants, potentially providing the population-scale coverage that current biobanks lack. The NIHR BioResource is another research resource made up of over 350,000 participants who have agreed to be contacted for studies, with a shared infrastructure for biological samples.

The UK LLC combines 20 smaller studies into a participant group of over 400,000 from across the four nations. Their unique combination of cohorts provides a useful illustration of how participant trust can be maintained through transparent processes that enable cross-cohort linkage.

These are purely research assets for pre-defined consented cohorts, useful for discovery and generating new findings.

Domain specific data research hubs

These data assets are less associated with a single geography and focused instead on developing access to the data and tools to address a particular anatomical, diagnostic modality or disease area. These initiatives and organisations prioritise the data and processes needed to deliver research, often collecting data that is not available elsewhere in the ecosystem.

The National Pathology Imaging Co-operative (NPIC) aims to develop a national digital pathology platform for the NHS which supports direct care through a single Picture Archiving Communications Service (PACS) system and research by developing a secure data environment that sits on top of this image archive. The live system hosts 29 petabytes of imaging data from ~40 hospitals of which 2 petabytes are available for research and linked to Genomics England.

The British Heart Foundation (BHF) Data Science Centre works with English, Welsh and Scottish TREs rather than developing their own. Their approach has been to use their expertise in data engineering and data science to develop scalable tools that support research in cardiovascular disease. Examples include a national data dashboard, reusable pipelines for analysis and software to support clinical cohort linkage.

Dementia Platform UK (DPUK) is built on top of SeRP infrastructure providing global access to over 100 multimodal data assets on 3.5 million individuals and supporting 36 ongoing studies and over 50 academic papers per year. DPUK is developing federated capability through partnership with the Alzheimer's Disease Data Initiative (ADDI).

Insight is an ophthalmic imaging bioresource holding over 30 million images across nearly 2 million patients based at Moorfields Eye Hospital NHS Foundation Trust. Through the Alzeye project they have linked routinely held data on patients with Alzheimer's to retinal images and used this to develop foundational AI models.

Place-based integrations

Over the last twenty years, there have been examples of local initiatives that have developed infrastructure and data that national initiatives have failed to achieve. Many of these initiatives have been maintained long after the initial development after demonstrating value to the population providing the data.

Connected Bradford, providing the only example of direct linkages between primary and secondary care EHR systems in England has been developed from the Born in Bradford programme (started in 2007), the global-first experimental birth cohort study to improve the lives of the population in Bradford.

The Salford Lung Study (2015) generated evidence of a new drug's effectiveness in a large, real-world patient cohort for the first time through integration with acute EHR systems. This enabled near real-time monitoring of trials and an evolution of this powers the NWEH service today.

The Whole System Integrated Care programme (WSIC) was developed in 2013 to improve the health of a 2 million population in North West London. It demonstrated that local data sharing agreements could be agreed and the integrated dataset that links primary and secondary care is still the most comprehensive population dataset in London.

NIHR Biomedical Research Centres (BRCs)

NIHR BRCs represent collaborations between NHS trusts and universities. The NIHR has allocated nearly £800 million to 20 BRCs across England. Their role is to bridge the gap between early-stage scientific breakthroughs and their translation into practical new treatments. Through these investments infrastructure has been developed that has demonstrated capabilities useful to an HDRS.

Data extraction for epidemiological research (Dexter) developed by the Birmingham BRC is software used to

automate epidemiological research within the West Midlands SDE.

Developed by Maudsley BRC, the Clinical Record Interactive Search (CRIS) system investigates mental and physical health through software that provides anonymous data insights.

The Francis Crick and Wellcome Sanger institutes

These institutes have been designed from the ground up to support biomedical research. Both have on-premise hardware and processes that are unique to their organisations but demonstrate capabilities that an HDRS requires.

The Francis Crick Institute houses around 1,700 people, including 1,500 researchers and 200–300 operational staff with high-performance compute and lab equipment on premise. They have developed the TRELLIS architecture, a cloud-based data fabric that combines governance and infrastructure as code with distributed computing capabilities. The Crick has a librarian-led data catalogue, archiving research that links back to datasets, data provenance and repeatable science stored for 20 years off-site.

The Wellcome Sanger Institute is primarily genomics focused, generating around 2 petabytes of data each year and 90 petabytes in total. They have their own on-premise network and computing structure. The technology bottleneck for their infrastructure is the throughput of data rather than compute, with limitations on reading and writing data more significant than vRAM of GPUs. The volume of data produced causes challenges for collaboration and limits the utility of cloud providers.

Disease registries and audits

Disease registries and audits span from national collections to small groups of academic clinicians. They are a critical resource for understanding the different data fields required for research for a given disease.

The National Disease Registration Service (NDRS) acts as the primary authority for disease registration in England, managing the collection and quality assurance of data concerning cancer (NCRAS), and congenital anomalies and rare diseases (NCARDRS). Other large national audits include the National Cardiac Audit Programme (NCAP) and the National Diabetes Audit (NDA), both of which are accessible through the NHSE SDE.

Outside of England there are longstanding cancer registries across the three nations; the Scottish Cancer Registry (SMR06), the Welsh Cancer Intelligence and Surveillance Unit (WCISU) and the Northern Ireland's Cancer Registry (NICR).

Hospital data paired to TREs/SDEs

Hospitals – particularly acute teaching hospitals, paediatric hospitals, and mental health hospitals – hold rich data unavailable in national extracts. These include detailed clinical documentation, diagnostic imaging, laboratory results, and specialty-specific datasets. Research access is provided through various arrangements: hospital-managed analytics environments, formal TRE partnerships with universities, or hosting within regional SDEs.

University-hosted TREs represent the traditional model, with academic institutions operating secure infrastructure (sometimes with HPC) that receive trust data flows under data-sharing agreements. These arrangements are particularly developed at academic health science centres. NIHR Biomedical Research Centres at 20 English sites[110] often anchor these partnerships. Increasingly, hospitals are moving to cloud data infrastructures that support data engineering, research, and deployment of their own data, outside a university TRE.

Hospital-level assets operate closer to source systems than regional or national infrastructure, with shorter extraction pathways and potential for richer data, including unstructured text and imaging. This proximity provides depth of data but limits coverage to single-hospital populations through bespoke arrangements. The landscape is fragmented with limited interoperability.

For the HDRS, hospital-level assets are essential for use cases that require depth beyond what national extracts can provide, but integration challenges are substantial.

Commercially owned assets (e.g., IQVIA, Arcturis, Flatiron, Akrivia)

A parallel ecosystem of commercially owned health data assets serves pharmaceutical and life sciences customers. IQVIA[111] aggregates data through commercial arrangements with GPs and hospitals. Flatiron Health[112] has built oncology-focused assets through cancer centre partnerships and by funding manual curation of detailed datasets that are usually inaccessible, offering commercial access through its own TRE. Arcturis[113] partners with hospitals to extract datasets for commercial real-world evidence. Akrivia Health[62] has developed NLP approaches to enrich mental health document-based records for real-world evidence and drug discovery. Commercial assets span all data-flow layers, with additional intellectual property derived from disease-specific pipeline development and AI data enrichment.

Commercial assets reflect market demand that public infrastructure has not fully met, particularly around access speed, curation depth, and service responsiveness. Their presence validates willingness-to-pay assumptions underlying HDRS commercial projections. If HDRS offers service levels comparable to those of the NHS, with comprehensive population coverage and authoritative provenance, it may capture market share. If access remains slow, commercial alternatives will continue attracting customers despite coverage limitations. Some commercial operators may be positioned to provide value-added services on top of HDRS infrastructure, rather than competing directly.

Federated research networks and tools

Federated research networks are an emerging infrastructure type that integrate distributed assets, enabling analysis or

AI model training without centralising person-level records. These networks therefore operate at the asset integration layer.

The DARE UK programme is currently developing technical components and demonstrator projects and exploring federated approaches, including work led by SeRP to establish cross-TRE federation capabilities[114]. They have developed a series of demonstrations which utilise Global Alliance for Genomics and Health (GA4GH) standards to describe a task execution service (Five Safes TES[158]) that could support complex orchestration of different analytical workloads, including machine learning. The programme defines the different analytical workloads in terms of whether the analysis can be completely isolated (traditional federation), whether it requires some connection (such as sharing weights – numerical parameters within neural networks – for an AI model) or whether it is centralised where row level data can be pooled.

Other tools occupy different positions in the federated landscape: DataSHIELD provides a software framework for federated analysis in the R programming language[115]. Vantage6[116] offers infrastructure for orchestrating federated learning across distributed nodes, and emerging tools such as Flower AI[117] and the Federated Learning and Interoperability Platform (FLIP)[118] address specific aspects of multimodal federated learning. Commercial platform providers such as Lifebit[72], Bitfount[73], Fitfile[74] and TriNetX[75] offer federated technologies and have partnerships across the UK ecosystem but none are providing national coverage.

Other infrastructure includes the NHS Federated Data Platform[8], which directly integrates with over 100 NHS trusts' systems to power operational use cases but which does not include provision for processing data for research.

The maturity of federated networks varies considerably, but in general, networks are yet to transition to stable, customer-ready assets in their own right. DataSHIELD has demonstrated utility in epidemiological research across international cohorts but is largely used in academic settings for ad hoc studies. The DARE UK programme has demonstrated technical feasibility for cross-TRE federation but platforms have not yet transitioned to production services capable of supporting researcher access to data networks. Flower AI and FLIP have demonstrated machine learning studies on distributed multimodal data assets, but are yet to form multi-regional networks.

In all cases, a substantial barrier is the availability of data at nodes with suitable compute, the correct data types, similar provenance, and common data models.

Summary

The UK health data landscape comprises substantial capability developed over decades, but this capability is distributed across asset types that largely operate in isolation.

Three observations warrant emphasis. First, considerable capability exists: the UK is not starting from a blank slate, and multiple assets have demonstrated research value and, in some cases, commercial viability, through capabilities that have been hard won.

Second, assets have largely developed independently, with limited evidence of integration or cross-asset analysis.

Third, the maturity, population, data coverage, and technical characteristics of assets vary substantially, and none constitutes a 'perfect' research asset for HDRS end-users.

A detailed assessment of functional and technical gaps – and how these constrain the user stories that HDRS must serve – is provided on page 39.

What capabilities are present?

The landscape description establishes what data exists, how it flows, and what assets have been built. This section applies the gap analysis methodology to evaluate current capability against the six core HDRS capabilities. For each capability, we assess what works and what does not. There is not one architecture or technology that can deliver all these capabilities. The gaps that exist today are largely due to the strategy taken by individual organisations to prioritise certain research needs over others. The findings from this gap analysis inform the subsequent technological opportunities.

Access to comprehensive longitudinal health records

Comprehensive health records mean different things for different user stories. Pharma real-world evidence studies require broad population coverage, cohort sizes, and pharmacy data with patient-level endpoints. Clinical trial sponsors need sufficient depth for eligibility assessment. Academic researchers may need historical data spanning decades. AI developers require diverse, representative data across demographic groups.

Strengths

Longitudinal primary care records with established quality assurance. For example, CPRD holds decades of GP records with documented provenance and established data quality processes. For real-world evidence studies requiring long-term medication exposure, disease natural history, or post-market surveillance, this represents mature capability, while a close relationship with contributing GP practices enables data quality validation against source systems. OpenSAFELY demonstrated 98%+ English GP coverage during COVID-19, proving population-scale primary care analysis is technically feasible when vendor cooperation and information governance permit.

Established secondary care administrative data. English Hospital Episode Statistics[119], Scottish Morbidity Records[120] and the Patient Episode Dataset for Wales[121] provides comprehensive coverage of admitted patient care, outpatient attendances, and A&E. For many users (for example, in Health Economics and Outcomes Research) that primarily require hospitalisation patterns, high-level diagnostic coding, or procedure volumes, these are mature and national assets that can be linked to primary care data where data sharing agreements are in place. These data assets have provided harmonised data for nationally representative analysis. Although primarily designed for commissioning services they have been utilised for epidemiological studies successfully.

Integrated datasets for regional and national populations. SAIL links primary care, administrative secondary care, prescribing, mortality, and non-healthcare data at the Welsh population scale[122]. DataLoch[100], Discover NOW[123], North West SDE, and Kent, Medway and Sussex SDE provide similar linked views for their populations[9].

Mortality linkage. The ONS mortality linkage is well established through multiple pathways[124]. User stories requiring treatment outcomes or survival analysis have viable data routes.

Challenges

The route to UK-wide GP coverage is unclear. Replicating pandemic-era coverage for routine research is not straightforward. Numerous national programmes have failed to achieve the requisite data controller and public assent. Primary care data sharing agreements remain a barrier for Scottish Safe Havens outside of DataLoch despite the technical pipelines already flowing data. The lack of legislation in Northern Ireland limits the research use cases for accessing centralised GP data.

Secondary care data, including prescribing information, is difficult to access and not easily linked. Hospital systems hold prescribing data, laboratory results, clinical documentation, and speciality-specific datasets that administrative data cannot capture. Prescribing data is a key gap repeatedly highlighted by stakeholders. A high-cost drugs dataset linking all 42 ICBs was developed as a proof of concept in April 2020 but has not been maintained as a national asset. Almost no sites demonstrate comprehensive linkage across the whole clinical pathway, including GP data. For user stories requiring understanding of hospital exposures and outcomes, this is a critical gap.

Assets, including those in four nations, are isolated from one another. Some data assets across the four nations and English SDEs contain comparable datasets that could be integrated to create a larger population representation. However, no deployed technology solutions can integrate these at scale.

Access to research-ready datasets from major research studies and biobanks

Research-ready datasets from consented cohort studies and biobanks provide deep, curated data on defined populations with explicit consent for research use. These

assets are essential for studies that require data beyond routine NHS collection, including genomics, imaging, lifestyle factors, and biological samples. Presently, precision medicine researchers requiring linked multiomics data (including international researchers) will go to UK Biobank and Genomics England. In the future, Our Future Health will provide a larger and more representative resource as more of its consented cohort are sequenced.

Strengths

World-leading cohort infrastructure exists. The UK Biobank, Genomics England, and the emerging Our Future Health represent globally recognised research resources that support thousands of studies a year. UK Biobank and Genomics England hold extensive multimodal datasets, including petabytes of genomics and imaging data.

Established research access models with clear commercial viability. UK Biobank's researcher access system is referenced as an example of a scaled research access service which delivers a reliable timeline and transparent pricing.

High compute TRE-based access infrastructure is operational. UK Biobank's Research Analysis Platform, built on DNAnexus infrastructure, provides researcher tooling, including an advanced data science and machine learning stack with access to higher compute requirements. Genomics England similarly operates TRE infrastructure with HPC capabilities. These represent mature, production-grade research environments.

A model for clinical integration exists. Genomics England demonstrates how genomic data can link to the NHS Genomic Medicine Service, enabling the return of findings to clinical practice.

Challenges

Linkage to routine NHS records remains difficult. Both UK Biobank and Genomics England report persistent challenges linking to primary and secondary care[109]. Despite consent, GP linkage has historically encountered governance barriers. Even mortality data presents recurrent quality issues. For precision medicine user stories requiring longitudinal outcome tracking, this limits value. On 10 February 2026 a data provision notice was published to enable linkage to English primary care data based on the GP Data for Pandemic Planning and Research (GDPPR) data specification. This includes coded data for diagnoses, prescriptions, referrals and lab results. This is a large step forwards for enabling linkage of cohorts within UK Biobank, Genomics England and Our Future Health, to routinely collected data.

Cohort populations are not representative. Volunteer bias means existing cohorts do not reflect the UK population[126,127]. UK Biobank participants are healthier and less diverse than the general population. Our Future Health aims to address this, but remains early-stage.

No mechanism exists for cross-cohort analysis. Each cohort operates its own TRE with separate access processes. There is no capability to query across cohorts. Rare disease research, requiring identification across multiple sources, is particularly constrained.

Access to high performance compute is often lacking. Discovery sciences are increasingly leveraging machine learning solutions with corresponding high compute requirements. Although cloud based TREs provide access to scalable GPUs, the cost of training and testing these models on the cloud exceeds what many researchers are accustomed to from university computing clusters. A number of studies in UK Biobank have expressed limitations from these constraints.

Access to advanced, multimodal diagnostic data

Advanced diagnostics data includes imaging, laboratory results, genomic data, and clinical text. For medtech and biotech developers building AI-based tools, access to representative, multimodal data is vital. Models trained on imaging, pathology, or clinical text require large volumes of diverse data to achieve clinical-grade performance and regulatory acceptance. The ability to move from development through validation to deployment in clinical systems depends on consistent access to data that reflects real-world clinical practice. Precision medicine researchers similarly require linkage between multiomics data and detailed clinical phenotyping to identify disease subtypes and validate biomarkers.

Strengths

Scotland has a national imaging research infrastructure. This national imaging archive has research access capabilities that cover all NHS boards, demonstrating that population-scale diagnostic imaging aggregation is technically achievable.

PACS infrastructure is universal. The underlying imaging data is stored at source across acute trusts, adhering to common standards. The challenge is extraction, quality assurance, and linkage.

Pockets of excellence at academic health science centres. Several major academic centres, particularly those with NIHR Biomedical Research Centre funding, have established imaging-to-research pipelines within their trust environments. These demonstrate that technical pathways from PACS to research-ready datasets can be built, though they remain institution-specific and dependent on local expertise.

Emerging capability for clinical text extraction. NLP tools for extracting structured information from clinical notes are maturing, and more organisations are seeing active deployments. There are early but meaningful steps toward unlocking unstructured data.

Northern Ireland has achieved high levels of digitisation across diagnostics, pathology, and cancer registration, including comprehensive population-level registries. However, extraction, standardisation, and governance

alignment vary by data stream, and some assets (e.g., pathology reports) remain easier to access as narrative outputs than as structured, reusable datasets. Investment is required to bridge the gap between digitisation for care and standardisation for research.

Challenges

Unstructured data remains largely untapped. Billions of clinical notes contain valuable information that remains locked in source systems. There is an infrastructural barrier to deploying information extraction tools in hospitals. A particular issue is unstructured primary care data, which presents both an information governance challenge and a challenge in obtaining such data from monopolistic GP EHR vendors.

Laboratory results are fragmented. Laboratory Information Management Systems (LIMS) vary by trust and even by department. There is no national laboratory data asset. Pathology results essential for phenotyping are largely inaccessible to research. These results are often provided back to EHR systems in a structured format that is not compatible with the data models in those systems. These are then stored as text fields rather than structured data and often automatically excluded from research processing pipelines as they are assumed to be free text and potentially disclosive. Stakeholders consistently identify access to this information as critical.

No national imaging asset exists outside Scotland. Imaging data remains siloed within individual trusts or SDEs, with no standardised extraction pathways, common formats, or national aggregation. For AI developers requiring diverse training data, this is a fundamental barrier. An additional challenge is sharing the information about these image registries through human readable catalogues. Meta data about the images (location on body, type of imaging, clinical observations) are not used consistently and so data discovery about what an image holds to share with researchers is challenging and requires manual checks.

High-performance computing is not co-located with data. In its current state, the UK supercomputing infrastructure is not designed for persistent, sensitive data storage. No large-scale NHS data studies have run on the UK supercomputer network. For user stories that require AI model training on large multimodal datasets, this separation is a material constraint.

No linked multimodal data at scale. Currently, the only realistic route to accessing advanced diagnostics data with longitudinal linkages is to work directly with hospitals and partnered universities who are willing to establish bespoke partnerships – a pathway that does not readily scale.

Data infrastructure is built for early development research, not for translation, where researchers build advanced AI models and biomarkers. Few sites have infrastructure in proximity to clinical systems that allow prospective deployment of AI models into live environments for validation. Without the ability to generate this type of evidence, it becomes harder to bring new innovative technology to market and funding is repeatedly funnelled to pilot initiatives.

Faster clinical trials and new treatment development

For data-driven clinical trials, speed and recruitment success are critical metrics. Sponsors need a rapid turnaround from protocol design to first patient enrolled. This capability is central to HDRS commercial projections and UK competitiveness for global trial investment. Clinical trial sponsors require integrated tools for feasibility, site selection, and patient identification. For precision medicine trials requiring biomarker-stratified recruitment, eligibility criteria cannot be assessed against administrative datasets alone. The current landscape largely fails to meet these expectations.

Strengths

NHS DigiTrials provides a national feasibility query capability. DigiTrials, operated by NHS England, enables trial sponsors to query aggregate patient counts against eligibility criteria across Hospital Episode Statistics and primary care prescribing data, for the entire population of England[95]. It interfaces with NHS digital communications infrastructure for direct patient outreach, providing a direct pathway to recruitment. It has enabled large trials such as RECOVERY during the COVID-19 pandemic[128], the Galleri Trial for early detection of cancer[129], and supported recruitment to Our Future Health. The software used to recruit and reach out to patients has demonstrated ability to scale, recruiting tens of thousands of people a week.

CPRD offers established feasibility and recruitment. CPRD facilitates patient case-finding and recruitment through a well-established GP network. The close relationship with practices enables recruitment without separate information governance approvals for each exercise and allows deeper feasibility and recruitment to be conducted with the support of clinical teams, increasing the conversion rate.

NWEH provides deeper clinical data. It enables near-real-time primary and secondary care data for enrolled patients enabling trials to monitor outcome data and make decisions about safety and effectiveness fast[96]. It also opens up the possibility for adaptive trials that leverage newer approaches for maximising the information from a trial using multi-armed approaches with varying the number of patients recruited to each arm. However, this requires deep integrations with acute systems which has limited it to regional populations where these are in place.

Challenges

DigiTrials remains constrained by pilot status, governance requirements, and distance from clinical teams. Under current pilot directions, recruitment support is capped at approximately 10 trials per year. Each recruitment exercise requires a Section 251 application, adding to timelines. Full-service directions are pending; until resolved, scaling to

meet commercial demand is difficult. Outreach is by mass communication rather than contact from clinical teams, and the service is less suited for trials that require a high recruitment success rate or those requiring clinical input at the recruitment stage as the eligibility must be based on nationally available data which is shallower, lagging and does not include linked primary care data.

Hospital-based and specialist trials are not served by national infrastructure. Most high-value trials occur in hospitals and specialist centres, particularly academic teaching hospitals and major clinical trials units, yet no national data-driven infrastructure exists to support these. Standard clinical trial processes in these settings remain manual, dependent on individual site capabilities and clinician relationships rather than systematic data access.

Invitation fatigue and population-level coordination are persistent risks. Without national-level visibility, individuals may be over-invited to trials. This is particularly problematic where data does not follow patients across systems. National coordination is needed to manage recruitment sensitively and avoid harm.

Cancer and precision medicine trials lack systematic support. Cancer data, whether in hospitals or registries, is not yet linked to recruitment capabilities. NDRS data is incomplete, especially for biomarkers and there can be 18 months of delay before records are available for analysis. There is no platform for biomarker-driven patient identification.

Feasibility queries are limited by data depth. Clinical records often lack industry-grade outcomes, are frequently missing key socio-demographic information, in general are broad and shallow, rather than deep, and often lack sufficient linkage for commercial trial precision[130]. This risks high pre-screening failures, delays in recruitment, and increases in cost[69].

Simpler access with a single entry point

A single entry point to disparate research environments with streamlined agreements and approval processes is primarily a service and governance challenge rather than a technology problem. However, the technology components discussed here underpin discovery and a consistent user experience across the distributed infrastructure. All six user stories require predictable, timely access. None can tolerate multi-year delays. Pharma customers specifically require transparent pricing and access timelines.

Strengths

HDR UK Innovation Gateway provides a UK-wide metadata catalogue. The Gateway catalogues dataset descriptions, variable lists, and access information across multiple data custodians, enabling researchers to understand available assets without direct access to the underlying data[131].

The Gateway cohort discovery tool enables federated queries across participating assets. Where implemented, researchers can query aggregate patient counts across multiple data assets to assess study feasibility. This represents a technical proof of concept for federated discovery[67].

Scotland's eDRIS provides a single national coordination point. eDRIS functions as a triage and routing service for Scottish health data access, demonstrating that national coordination across assets is achievable within a devolved nation's governance framework[132].

The Honest Broker Service in Northern Ireland has been running an access service since 2014. They demonstrate a high degree of transparency with detailed reports on project access, datasets used, outputs checked and modifications to any projects.

Challenges

HDR UK Gateway coverage is incomplete, and metadata quality is inconsistent. Many UK data assets are not catalogued or have outdated metadata of varying granularity and usefulness. The Gateway is dependent on voluntary asset submissions. Stakeholders note that current metadata does not provide sufficient insight into whether data is 'fit for purpose', nor an indication of whether downstream processes for gaining access are functional or reliable.

Cohort discovery requires substantial investment at each node. Federated cohort queries require CDM transformation, infrastructure setup, and ongoing technical maintenance across participating assets. Few assets have completed this work. The approach is inflexible to complex feasibility queries that require data outside of the CDM.

Governance and contracting timelines are slow and unpredictable. This was a common theme and has been highlighted in previous reviews[5]. Access processes operate through sequential approvals with limited visibility into status or expected duration. Stakeholders report that identical requests to different controllers yield different outcomes and timescales.

TRE user experience varies substantially. Software availability, compute capability, and interface design differ across environments. Stakeholders report arbitrary rules governing available tools, inconsistent package mirrors, and slow responsiveness. For researchers working across multiple TREs, this creates friction and project risk.

Linking multi-sector data for greater impact

Linking health data with information from other sectors, such as environmental, social, education, and housing, enables research into determinants of health beyond healthcare delivery. This capability requires both technical linkage mechanisms and governance frameworks permitting cross-sector data sharing whilst maintaining public trust. Typically, academic researchers and public health bodies may value broad linkages to understand social and environmental determinants of health, and the ability to translate insights into real-world pathways that target vulnerable communities and individuals.

Strengths

SAIL demonstrates comprehensive health-to-non-health linkage at the population scale. SAIL has achieved person-level linkage to schools, education, housing, justice, census, HMRC, social care, and environmental data for the Welsh population[99]. This represents the most comprehensive cross-sector linkage capability in the UK and includes technical capability for probabilistic linkage of datasets without NHS numbers.

ONS Secure Research Service (SRS) holds linked administrative data across government departments. The SRS provides access to linked census, tax, benefits, and other administrative data for approved research projects[133]. While not health-focused, it demonstrates that cross-departmental data sharing is achievable within appropriate governance frameworks.

Administrative Data Research UK (ADR UK) is a four-nation effort to link administrative data in a secure manner that preserves citizens' privacy. It has over 14 government departments engaged in research projects across 350 datasets. The programme has had £168m confirmed funding over next five years with a recent evaluation estimating that every £1 of investment returns £5 of public value.

Challenges

No equivalent cross-sector linkage capability exists in England, Scotland, or Northern Ireland, at a population scale. Despite England holding the largest health datasets, systematic linkage to education, housing, benefits, employment, or social care data has not been achieved. The Scottish Government's Care Reform Scotland Act provides a legal mandate to integrate social care records with health data[134], but comprehensive cross-sector linkage has not yet been demonstrated at scale.

No standardised approach to cross-sector linkage. Non-health sectors use different identifiers (National Insurance Number, Unique Pupil Number, etc.). Person-level linkage across sectors requires either a universal identifier, which does not exist, or probabilistic matching using demographic attributes[135]. While SAIL has shown success in the latter, there is no standard approach, and results are clearly inconsistent across assets[137].

Social care data suffers from poor data quality and fragmentation. Social care is delivered by local authorities who capture data using heterogeneous systems with inconsistent standards[136]. Access to this data is a governance challenge, and linkage to other datasets cannot rely on canonical identifiers. Analytical value is limited by poor data quality and low completeness.

Inequalities are exacerbated by lack of representation for key, vulnerable groups. Data on marginalised populations, such as children in the care system, are not routinely linkable[139]. This is often due to complex governance around sensitivity of such data, but such exclusion risks deepening health inequalities[138]. The system lacks fair, proportionate pathways to access this data to ensure vulnerable groups are not invisible in the national data landscape.

What technologies are missing?

The gap analysis identified where current capability falls short. This section translates those gaps into technical requirements, organised by data flow layer.

Source systems layer

Sufficient source system extraction pathways do not exist

The UK's clinical data originates in dozens of heterogeneous source systems. GP clinical systems (EMIS, TPP, Vision) have established extraction pathways used by multiple downstream assets, but access terms, pricing, and cooperation vary by vendor. Hospital EHR systems pose a significant challenge, as the full richness of back-end data systems is difficult to access. An example of the complexity observed in a single trust is found in Table 3..

Developing integrations with these systems takes months of work and thousands of lines of code, and the resulting code may be sensitive to software updates or changes in upstream data models.

HDRS opportunity: A pragmatic strategy that does not attempt to build all source system connections at once, but identifies priority extraction pathways for high-value data types (e.g., hospital prescribing and drug administration, laboratory results, imaging, biomarkers in clinical text).

Table 3. An example of the number of systems in a single trust

Clinical area	System
Bloods	BloodTrack
Maternity	BadgerNet
Observations	Cerner
Medications	Cerner
Procedures	Cerner
Diagnosis	Cerner
Cancer tracking	Infoflex
Cancer therapy	Aria
Radiotherapy	Aria
ICU	CareVue
Endoscopy	Endobase
Colorectal	Janus
ECG	Schiller
Echo	ISCV
Cardiovascular	Solus
Pathology	CellPath
Radiology	CRIS
Genetics	STARLIMS
Labs	WinPath

Source data processing layer

Production-grade data pipeline infrastructure is scarce

Mature data assets rest on mature pipelines. CPRD and SAIL have invested years in building robust, documented transformation logic with quality assurance processes. This capability is rare. Most NHS organisations lack the infrastructure (and expertise) to build, host, and maintain production-grade pipelines that can 'clean' data into more analysable structures, while providing data quality understanding and monitoring. Unstructured data processing is a significant capability gap, with primary care free text particularly constrained.

HDRS opportunity: Scaling up of pipeline infrastructure to acceptable standards, and widespread deployment of Language AI tools to unlock unstructured data assets.

Data asset layer

No mechanism resolves a patient to their complete set of health records

For a comprehensive longitudinal record to exist, there must first be a mechanism to identify what data has been generated for a given individual across the health system. The NHS number provides a foundation, but the infrastructure to link a person to their complete set of health records – across primary care, secondary care, speciality systems, registries, and research datasets – does not exist in any comprehensive form. NHS England operates a Master Person Service, but it covers only a subset of national datasets.

Without this capability, linkage becomes a series of bilateral negotiations between specific assets. Researchers cannot know what records exist for a cohort without separately approaching each source. Clinical trials cannot identify complete patient histories across care settings.

HDRS opportunity: A patient dataset resolution service that can identify, for consented or non-consented by legally permitted purposes, what data sources hold records for a given individual. This does not require centralising data, but

does require participating assets to maintain and expose patient indices against a common identifier, with appropriate governance controls.

Specialist and disease-specific data sources lack linkage pathways to routine records

Routine NHS data provides breadth but often lacks items that may be specially curated or recorded outside routine processes. Disease registries, biobanks, and research cohorts hold curated information – genomic sequences, detailed phenotyping, biological samples – that cannot be derived from routine extracts. These sources operate under different governance models with distinct identifier coverage, data structures, and access arrangements.

Linking specialist sources to routine NHS records is where substantial research value lies. UK Biobank and Genomics England report persistent difficulties achieving timely, complete linkages to primary and secondary care. Disease registries often cannot be linked to the clinical pathways that generated their entries. Technical capability exists, but governance complexity, inconsistent identifiers, and the absence of standard agreements create friction.

HDRS opportunity: Standardised governance pathways and technical interfaces for linking specialist data sources (including disease registries, biobanks, and research cohorts) to routine NHS records. This includes technical specifications for how specialist sources should expose linkage keys and metadata.

Linkage quality is unmeasured and unreliable

Even where pathways to link datasets for an individual patient are available, the resulting quality of linkage is a critical failure point. Stakeholders identify NHS England record linkage as poor quality and reliant on outdated algorithms, with some populations systematically missing[50] – a problem exacerbated by a lack of data standardisation. The absence of standardised metrics for reporting (match rates, false positive/negative rates) means researchers cannot assess whether their research findings are biased by linkage failures. Every downstream use then inherits linkage errors that users cannot quantify.

HDRS opportunity: Standardised, validated linkage algorithms with published performance metrics and ongoing quality monitoring.

Minimum data specifications do not exist

There is no agreed definition of what data types a research asset should contain. Assets vary in whether they include hospital prescribing, laboratory results, cancer registry data, imaging, or clinical text. Otherwise detailed assets may lack entire information categories essential for specific user stories.

Beyond scope, technical standards are inconsistent. The same clinical concept may be represented differently across assets. UK vocabularies (dm+d, SNOMED CT subsets) are applied inconsistently. Provenance recording and other documentation ranges from comprehensive metadata and dictionaries to being entirely absent. Quality metrics are unstandardised and usually missing.

HDRS opportunity: Minimum data specifications addressing scope (required data types) and technical standards (common data models, coding vocabularies, documentation, quality metrics). Published criteria enabling researchers to assess fitness for purpose before committing to access applications.

No data versioning, archiving, or reproducibility

Research reproducibility is typically framed around analytical code: can another researcher run the same analysis and obtain the same results? But reproducibility also depends on data stability. Health data assets are not static. New extracts arrive, linkages are updated, errors are corrected, coding schemes evolve, and populations shift. An analysis run today against a data asset may produce different results from the same code run six months later because the underlying data has changed.

Few UK assets offer versioned, point-in-time snapshots allowing reproduction against data as it existed at original execution[5]. Many operate rolling extracts where historical states are overwritten. Research outputs cannot be validated, regulatory submissions cannot be audited, and scientific claims rest on datasets that no longer exist.

HDRS opportunity: Data versioning and archival standards for participating assets, enabling point-in-time dataset snapshots to be preserved and accessed for reproduction purposes. This includes both source data assets and researcher-generated derived datasets, with retention periods aligned to research and regulatory requirements.

Asset integration layer

Asset integration infrastructure does not exist at scale

This layer – where separate assets are joined to enable analysis across larger populations – is where UK capability is weakest relative to HDRS ambition. Regional SDEs in England operate as isolated environments with no standardised cross-SDE analysis mechanism. Scottish Safe Havens lack automated federation, with cross-regional projects relying on manual coordination. Welsh and Northern Irish infrastructure, while more integrated internally, has no systematic connection to English or Scottish assets. Regional boundaries remain data boundaries.

Federated analytics has been demonstrated technically but not operationally. Barriers are partly infrastructural: investment required for node compatibility, lack of standardisation agreements, rigidity of predefined models, and also related to inflexibility in source data requirements and user reluctance toward 'eyes-off' approaches that restrict data exploration.

Data fabric architecture – unified discovery, access, and analysis through metadata-driven orchestration – does not yet exist at scale in the UK. Elements are emerging across some assets, and fabric patterns dominate large-scale data

management elsewhere, but these have not yet been integrated across assets.

HDRS opportunity: Prioritise the integration layer as a core technical contribution. This includes interoperability standards that participating assets must meet, the development of a common technology pattern, and governance frameworks that enable cross-asset queries on a common legal basis.

Research access layer

TRE capability, usability, and compute availability are inconsistent

TRE quality varies substantially. Some environments offer modern interfaces, flexible software, and scalable compute. Others provide constrained desktops with fixed software tools that do not support bringing in code, data or models.

High-performance computing is not designed for persistent sensitive data or for managing the research project lifecycle. No large-scale NHS data studies have run on UK supercomputer infrastructure.

Translation infrastructure is absent. No systematic pathway exists to prospectively validate AI models or biomarkers against live clinical data, or to deploy them in care settings. The gap between the research environment and clinical system means innovations cannot generate real-world evidence required for regulatory approval.

HDRS opportunity: Minimum TRE standards addressing usability alongside security, flexibility for compute co-location, sandboxed environments adjacent to clinical systems, enabling prospective validation pathways.

Service provision layer

Discovery metadata is insufficient for feasibility assessment

Data asset discovery services exist, but coverage is incomplete, metadata quality varies, and the information provided is insufficient for users to assess data fitness for purpose. The ability to query data assets to assess study feasibility or select assets (and sites) for research is largely missing. Clinical trial infrastructure highlights the gap between capability and readiness for high-value use cases.

HDRS opportunity: Comprehensive, high-quality metadata across all participating assets, with sufficient detail to support study feasibility assessment without requiring full data access.

High-fidelity synthetic data is not routinely available

Researchers often cannot assess whether a data asset is suitable for their study without first seeing its structure, coverage, and quality. Accessing data often requires completing access applications that can take months. Synthetic data offers a solution in which artificially generated datasets preserve the statistical, structural, and semantic properties of the underlying data asset[140]. UK assets such as OpenSAFELY have recognised this, although it is a mandatory requirement due to the platform's 'eyes-off' code-to-data nature. CPRD publishes synthetic data to support study design. These remain exceptions.

HDRS opportunity: Synthetic data that represents data asset formats and structures, as standard capability for participating assets, with published generation standards preserving analytical utility while maintaining privacy protection.

Research workflow orchestration does not exist

A research project using health data must follow a sequence of activities, including identifying relevant assets, submitting applications, obtaining approvals, accessing environments, conducting analyses, exporting outputs, and ensuring that all intermediate steps are archived and/or reproducible. Currently, each step is managed through separate systems across separate organisations with no integration or standardisation. Progress is most often tracked through emails, conversations, and spreadsheets.

This fragmentation creates inefficiency, duplication, and risk, particularly at handoff points. For example, approvals granted in one system are not visible in another. Data and analytical provenance are not carried through to final outputs, and any retrospective review (for example, by a regulator during a Health Technology Assessment) becomes an exercise in archaeology.

HDRS opportunity: Digital workflow orchestration providing end-to-end visibility across participating assets, including standardised project tracking, and comprehensive provenance capture enabling reproducibility.

What are the constraints?

Technology does not exist in a vacuum; the UK health data ecosystem operates as a complex sociotechnical system where technical infrastructure is a necessary but insufficient condition for success. Our analysis identifies non-technical constraints clustering into five interconnected areas: governance and accountability; commercial and market-making; stakeholder engagement and values; operational capacity; and research coordination.

Governance and accountability

Well-designed technical infrastructure can protect privacy and enable transparency, but cannot ensure ethical justifiability or social acceptability. The governance framework supporting health data research is therefore as important as the technology it supports.

Information governance fragmentation

The shift toward SDEs was intended to simplify governance by standardising privacy-by-design[5], yet fragmentation remains high. Each SDE operates its own strategy, while approval bodies (Confidentiality Advisory Group (CAG), Research Ethics Committees, Data Access Committees, Patient Advisory Groups) remain structured around project-specific approvals rather than pan-network access. This creates a 'detective work' burden for researchers navigating multiple controllers. Requests through a single front door, such as the Data Access Research Service, typically take six months or more; one study reported 2.5 years from setup to data acquisition[141]. Governance typically operates sequentially rather than in parallel: meaning a Data Access Committee can reject a study after it has received ethics and CAG approvals, leaving funded research undeliverable. Cross-jurisdictional work adds complexity, as devolved nations operate under different legal frameworks.

The cumulative effect undermines UK competitiveness. Developments such as the EHDS risk eroding the advantage conferred by UK data assets, while unpredictable data-acquisition costs pose barriers for start-ups and impediments to large-scale trials for established pharmaceutical companies[142].

Consent and the National Data Opt-Out

Health data research operates under two distinct consent models. For routinely collected data, the system uses an opt-out basis through the National Data Opt-Out, allowing individuals to prevent their records from being shared for research and planning[143]. Approximately 5.4% of the population has opted out, with opt-out rates spiking during trust crises such as the 2021 GDPR programme, likely undermining dataset representativeness[144]. There is a need to protect national data assets from further related harm in the future[145].

For specifically consented research – cohorts, registries, clinical trials – individuals explicitly opt in[146]. Yet patient consent does not guarantee data access. Controllership rests with healthcare providers, typically GPs, who make independent decisions. Some decline to share records even when patients have explicitly consented, requiring researchers to seek permission from each controller separately[5].

Addressing these barriers requires care. Public trust remains a foundational asset: surveys and stakeholder feedback consistently indicate broad support for using health data for public benefit, alongside frustration with processes perceived as unnecessarily complex or duplicative[147,148]. At the same time, the existing system of checks and approvals plays a critical role in ensuring privacy, purpose limitation, and accountability at each stage of data use.

The HDRS will need to navigate this tension deliberately – reducing avoidable burden and inconsistency while preserving the safeguards that underpin public confidence. Stakeholders highlighted heightened sensitivity around who has access to data and concerns about potential impacts on benefits or entitlements, alongside strong support for research that delivers clear and tangible patient benefits, especially in rare disease and paediatric contexts[146]. This reinforces the need for locally grounded consent models, transparent auditing, and explicit articulation of public value as integral components of any future HDRS approach.

Commercial and market-making constraints

While the UK maintains world-class health data assets, the current landscape requires greater structural cohesion to realise its full economic and clinical potential. At present, high operational friction and inconsistent pricing models create hurdles that can inadvertently direct investment toward international markets.

Industry stakeholders often describe the UK research environment as "unreliable and unpredictable." This perception has damaged the UK's global standing, shown by its drop in Phase III clinical trial rankings from fourth in 2017 to eighth in 2023[27]. The UK is the second slowest among 18 European countries for trial setup, while competitors like Spain have expedited regulatory targets and consistently achieve top rankings.

Industry partners increasingly prioritise speed and predictability in data access. Moving away from the current 'bespoke burden', where partners navigate over 100 individual negotiations with different trusts, is a critical priority. To compete effectively, the UK must transition toward unified frameworks, such as Master Services Agreements, to reduce legal overhead and move toward standardised 'time-to-data' targets.

Furthermore, a sustainable data economy requires a balanced value exchange. Currently, local data controllers often manage the legal risks and operational costs of data curation without a clear mechanism for reinvestment. By establishing a transparent revenue-sharing framework, we can ensure that local trusts are fairly compensated, ensuring that national research is viewed as a complement to, rather than a distraction from, local care delivery.

Ultimately, without an evolved service layer, there is a risk that industry will continue to seek more agile solutions from commercial aggregators. To ensure the NHS remains the primary partner for health innovation, we must modernise our service levels to match the speed and ease of the private sector, securing the vital revenue streams needed to maintain and enhance the UK's public health data infrastructure for the long term.

Stakeholder engagement and values

Health data is intensely personal. While a SNOMED code may appear to a researcher as a data point, to an individual, it may represent a diagnosis that changed their life, the loss of a loved one, or a condition never disclosed to family[5]. Routinely collected NHS data is also a public asset, generated through taxpayer funding and interactions with a healthcare system built on solidarity and universal access[147]. Accessing this data for research is a privilege that must be earned.

Shared mission

During COVID-19, data-driven research gained widespread acceptance because its public benefits were immediate, tangible, and universally relevant. In contrast, Care.data's six aims created confusion about whose interests were served. When the public cannot easily understand and endorse the core purpose, trust becomes fragile[149,150]. Large-scale health data research initiatives require a compelling shared mission – whether focused on specific high-burden conditions, reducing health inequalities, or accelerating treatments for unmet need. What matters is that the mission is developed with patients and the public, regularly evaluated, transparently reported, and demonstrably pursued.

Value alignment

Patients and the public broadly support the use of health data for research but are uneasy about private-sector involvement and perceived commercial exploitation[143,147]. This is fundamentally about value alignment: support stems from endorsement of the NHS's core values of solidarity, universality, and equality. Using NHS data solely to drive economic growth or generate profit for private companies, conflicts with this ethos. Benefits sharing – such as preferential NHS access to products developed using its data, or demonstrable improvements to data assets – can mitigate this tension. Consistent ethical review confirming alignment with these values is equally important.

Transparency

Patients and the public want to know who accesses their data, what is accessed, how it is used, where it is stored, when access occurs, and why[151]. Transparency must be built into health data research by default, through public-facing dashboards, searchable registries of approved projects, and feedback mechanisms that demonstrably influence decisions.

Co-design

True support comes from partnership. PPIE processes should ensure those affected by data use have meaningful input into how systems are built and governed[152]. Done responsibly, co-design builds trust[153]. Done poorly, it backfires[154] – making it essential that PPIE is taken seriously.

Operational capacity

Technical systems require skilled workforces and stable funding to bridge the gap between research prototypes and production-grade infrastructure.

Workforce and skills gaps

Specialist staff spanning data engineering, information governance, research management, and domain knowledge remain scarce and unevenly distributed[155]. Data transformations depend on asset-specific structures, requiring professional engineers with expertise and experience delivering technical solutions into production environments. These skills do not always overlap with those found in research and academia. Some organisations demonstrate superior capability, but this knowledge is rare.

Information governance specialists present a distinct bottleneck. Staff who can navigate both the commercial and legal dimensions of data-sharing agreements while preserving privacy are scarce.

These workforce gaps are difficult to close. Automation offers limited relief: nervousness persists about automated approaches, particularly for output checking, and the combination of technical, legal, and domain skills required resists easy codification. At the same time, there is no sector-wide coordination on training or workforce planning.

Funding

Funding instability constrains long-term development[156]. The stop-start nature of project-based funding, as well as grant-based academic funding, creates deep structural problems. Academic incentives favour publication and novelty over refactoring a codebase and scaling production software. Once a paper is published, the underlying code may accumulate technical debt with no budget for ongoing engineering.

This instability also makes institutional knowledge fragile. Researchers and group leaders at some institutions are limited to fixed-term appointments, and when the original data generators depart, understanding of the datasets erodes, complicating reuse and interpretation for years afterwards. The development of a dataset is also not a one time exercise. Resources are required to monitor and maintain pipelines as the underlying data changes so that new schemas are included and deprecated data removed. Short-term funding further hampers the adoption of modern computational infrastructure: cloud computing shifts costs from capital to revenue expenditure, and managing cloud costs, decommissioning legacy pipelines, and maintaining cyber resilience add ongoing operational burdens that are routinely underestimated.

In the absence of a stable public-sector mechanism for industrial maintenance, including enforceable service-level agreements and continuous reliability, academic tools cannot function as primary national infrastructure without considerable risk once specific grants conclude. This issue is compounded by the lack of professional career pathways for Research Software Engineers within the NHS or in universities, resulting in critical infrastructure being maintained informally by researchers and creating significant single points of failure[5].

Together, these pressures make it difficult to bridge the gap between research prototype and production-grade infrastructure. Successful tools remain fragile academic projects rather than robust, maintained assets.

Challenges in procurement and technological fragmentation

The methodology for purchasing technology within the ecosystem often creates organisational silos that actively resist national integration.

Current procurement practices frequently result in the adoption of analytics solutions that lack transparent data processing, thereby impeding research efforts that must comply with regulatory standards established by the MHRA or NICE.

Local purchasing decisions typically prioritise immediate clinical requirements rather than broader research interoperability objectives. Consequently, essential data is often confined within proprietary systems, and vendors lack incentives to facilitate data sharing. Although the Procurement Act 2023 permits more flexible and modular procurement strategies, local teams often lack the capacity to enforce open standards[157]. This situation commonly results in contracts that allocate intellectual property and code ownership to vendors, rather than retaining these assets within the health system.

Research coordination

The ecosystem's value ultimately depends on its capacity to generate high-quality, impactful research that translates into improved health outcomes. This requires coordination mechanisms that avoid duplication, enable collaboration across institutions, support reproducibility, and allow successful solutions to benefit the whole system.

Research management and communication

There is no single research management system across delivery networks, making it often unclear which projects are being conducted, where, and by whom. This limits efficiency, increases duplication risk, makes it difficult for the public to see how their data is used, and undermines the ability to track outcomes.

When the ecosystem conducts research, it often provides limited information about impact back to patients and the healthcare system. Patients are rarely informed about how their data changed knowledge or outcomes. There is no systematic monitoring of research impact on population health, meaning effects on equality are often missed. This slows translation and erodes trust.

Collaboration and openness

Collaboration patterns are institutionally concentrated: a disproportionate share is held by a small number of universities, shaping how platforms are designed and potentially limiting broader interoperability.

Tension also persists between open data ambitions and researcher incentives. Institutions support data availability statements, but researchers often remain protective of their data prior to publication to realise intellectual property value[5]. This friction cannot be resolved by policy alone.

Summary

These five interconnected areas (governance, commercial and market-making constraints, stakeholder engagement and values, operational capacity, and research coordination) function as a complex sociotechnical web where failure in one domain inevitably erodes progress in others. For example, fragmented information governance directly overburdens an already stretched and scarce specialist workforce. Simultaneously, persistent funding instability prevents the development of professional-grade tools that could automate processes and reduce that governance burden. Poor research coordination further compounds these issues, ensuring that technical solutions developed in one region remain unavailable to others facing identical challenges, which results in widespread duplication of effort.

Ultimately, failures across any of these pillars can erode the fragile public trust and social licence upon which the entire health data ecosystem depends. While technology provides essential levers to address specific bottlenecks, it cannot resolve these structural constraints in isolation; the proposed HDRS recommendations and pilot initiatives are designed to address this multifaceted reality.

From gaps to opportunities

The options evaluated in the opportunity analysis have resulted from a rigorous synthesis of existing evidence, services, and technology, current market valuations, and deep stakeholder engagement. First, we highlight the most critical technology and system gaps identified in the review, then connect existing system strengths and solutions emerging from the evidence that are directly relevant to bridging them.

Opportunities presented reflect options that are technologically feasible, viable given known constraints, and valuable for the research ecosystem. Illustrative pilot initiatives are presented on page 52 that have the potential to enable core technical features as dependencies, for delivering new capabilities to end-users of the HDRS.

Transitioning from national assets to national service

The landscape review demonstrates a 'federation of fragments', comprising individual assets of varying quality, without the necessary level of integration or scale. Previous attempts which focused on centralisation of all data into a single repository have encountered significant technical and social challenges.

This analysis emphasises a transition from a manual 'trust and contracts' model across individual assets to a 'technology guarantee' model administered by a central body, in which technology enhances data value for researchers, and guarantees their overall experience.

Technological gaps and opportunities

The opportunities outlined here emerge from the evidence and engagement conducted as part of this review, and are intended to demonstrate a range of improvements that are achievable and would result in significant impact. They are not intended as blueprints for implementation, and different combinations of opportunities may be appropriate depending on how the HDRS role and mandate develop over time.

Gap 1: Depth and consistency of data assets

Existing data assets often lack the depth and consistency required for high value commercial and research use cases. The same data types are represented inconsistently across assets, and provenance is often poorly understood.

There is not a definition of data types that an asset should contain. Valuable data types are missing from assets. There are scarce linkages between GP data and deeper hospital data. Standard vocabularies (dm+d, SNOMED CT subsets) are applied variably and rarely present in source data. There has not, to date, been available funding to bring source structures toward a standard.

Why is bridging this gap important for users?

Useful data is either unavailable (locked in source systems) or not usable together (incompatible representations).

Pharmaceutical real-world evidence studies cannot access hospital prescribing where specialist medications are initiated. Oncology trial sponsors cannot systematically identify patients by staging, therapy lines, and biomarkers. Precision medicine researchers requiring linked laboratory values must fund manual chart reviews or work outside the UK. Regulators cannot reliably obtain exposure counts across assets when clinical concepts are coded inconsistently or not at all.

<u>Opportunity 1: Defining minimum information standards across all four nations to fuel data quality improvement and high value linkage opportunities, and investing in their implementation</u>

Substantial capability already exists, representing decades of investment and established data controller relationships. There is opportunity to raise assets to a common baseline rather than replace them, while opening valuable data flows that do not currently exist.

Existing data assets across the four nations can be enhanced, prioritising high-value items such as hospital prescribing data, laboratory and pathology data, concepts extracted from unstructured clinical text, and linkages to primary care data. This could be delivered through defining minimum information standards for participation in the HDRS.

Why this matters: High-value clinical data (hospital prescribing, lab results, clinical text) are locked inside NHS source systems, and unavailable to the research service.

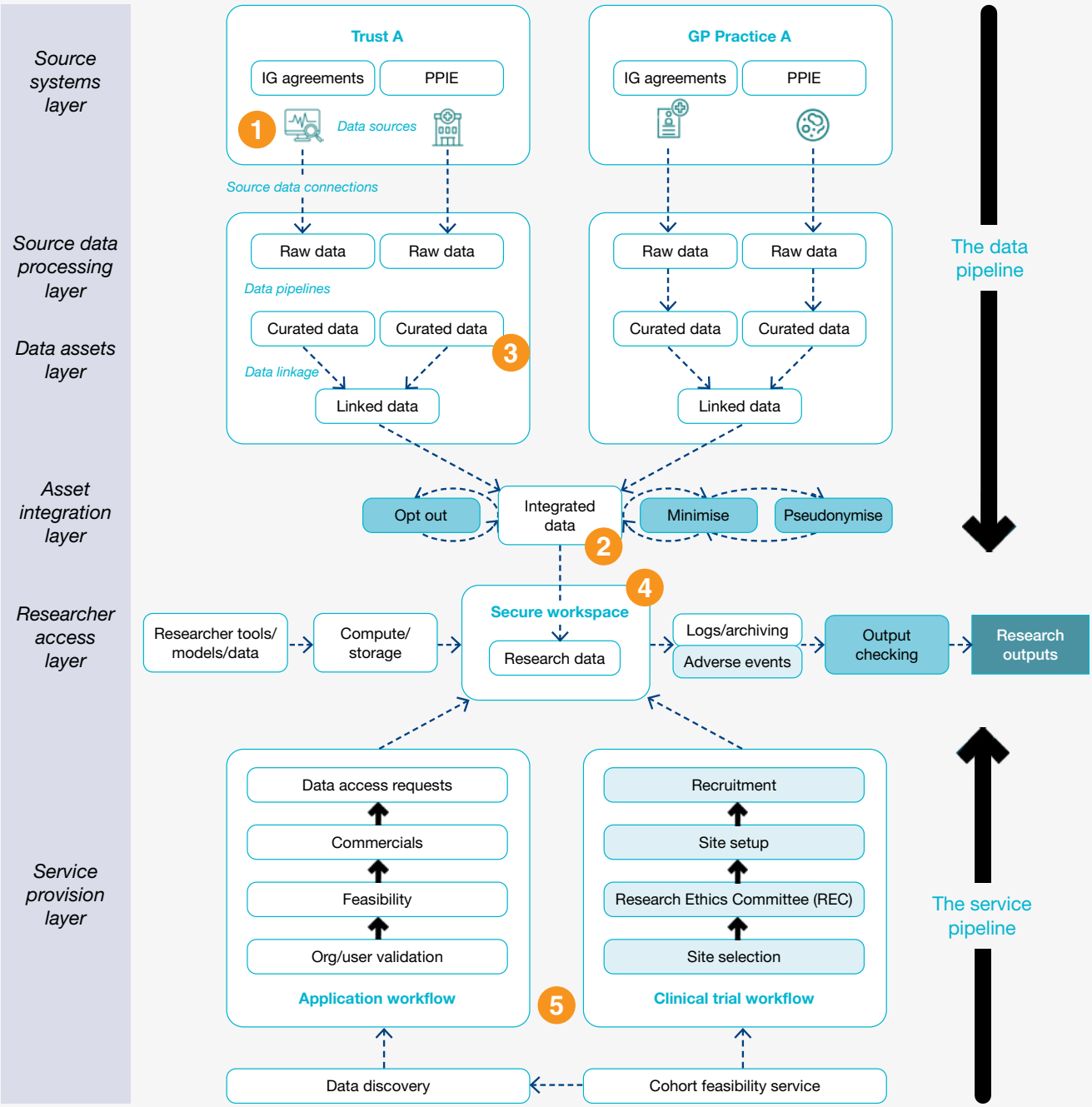
How this brings value: Enhancing data assets in the same way produces a consistent national data offering that can credibly compete for commercial contracts.

What could this look like?

A standard could specify two things: (1) the types of data and linkages that assets must contain, and (2) how that data must be represented.

For data types, targeting hospital prescribing, laboratory and pathology results, linked where possible to primary care data for contiguous populations, could achieve value for HDRS end-users.

Figure 6. Five technology gaps identified across the six layers



Key

Components | Processes | Clinical trials related | Technological gaps/opportunities

Technology gaps as mapped to components of research data infrastructure. These include (1) lack of depth and consistency of data flowing to data assets; (2) No scalable integration architecture to join data across assets and environments; (3) No standard mechanism nor reliable service for linking datasets for the same patient; (4) Variability in Trusted Research Environment usability and capabilities; (5) Slow and unpredictable access as key barrier to research.

For data representation, a standard could define semantic standards (nationally recognised coding systems with documented mappings), essential fields (minimum fields per data type ensuring consistent analytical use), and provenance records (source system, extraction method, transformation logic).

One option is to mandate transformation to a common data model such as OMOP. However, while this is valuable for specific applications, full transformation imposes high startup cost, rigidity, and potential information loss. The HDRS will need to consider whether it is necessary to impose this level of standardisation as a prerequisite for participation.

What value would this bring to the HDRS?

In the short term, minimum information standards would create clear accreditation criteria. Targeted investment in priority data types unblocks the highest-value use cases and provides transparency to users about what they can expect.

In the long term, engineering at source eliminates repeated per-project transformation costs and reduces time to data. Representation requirements ensure assets meeting the specification can be used together without extensive harmonisation.

This is an opportunity whereby the HDRS could help the UK transition towards data assets with a consistent national capability.

What is the reasoning behind this opportunity?

Stakeholder interviews consistently identified hospital prescribing, laboratory results, and items from unstructured text as most frequently required but least available, and source system extraction was identified as a critical bottleneck.

To date OMOP adoption has been uncoordinated, with organisations mapping similar data differently. While OMOP is valuable for observational research and regulatory uses, the consensus on imaging, genomics, cancer, and multimodal data is limited. Requiring semantic standards and provenance documentation enables harmonisation and easier analysis, even without mandating OMOP as a prerequisite.

Technical workshops validated that extraction and transformation capability is unevenly distributed and built to local specifications rather than national standards. Participants emphasised that curation should move upstream to maintain provenance understanding.

What are the risks of this approach?

Standard requirements may create thresholds some assets may not meet, and restrict certain use-cases. Specifications must accommodate legitimate variation while prioritising near-term value delivery.

Misalignment with emerging Single Patient Record specifications is possible. Early coordination would ensure research-ready data could eventually derive from an authoritative clinical account while remaining independent of SPR timelines.

Investment in data engineering at source requires sustained funding and technical capacity that may not exist in all regions. Building this capability in a competitive market may be slow and expensive.

Free text analysis at scale remains technically challenging. Current capabilities are largely disease-specific or site-specific. Bottlenecks are often infrastructural and cultural.

Coordinating standards across four nations introduces governance complexity given devolved arrangements, different legal frameworks, and existing investments.

What are the risks of prioritising other opportunities?

Without centrally defined requirements covering both data types and representation, HDRS becomes a coordination layer over assets of variable quality and incompatible structure. High-value use cases requiring medication exposure, biomarker validation, or granular phenotyping continue relocating outside the UK. Cross-asset analysis requires extensive per-project harmonisation.

The alternative is accepting existing asset variation and focusing HDRS investment on discovery, access, and governance. This may deliver faster initial progress but perpetuates structural weaknesses limiting UK competitiveness.

If pursued, what are possible steps?

Establish a minimum information specification defining required data types and linkages, accepted semantic standards, essential fields, and provenance documentation requirements.

Conduct a readiness assessment of existing assets against the standard, informing a phased investment plan prioritising assets with greatest potential to unlock high-value use cases.

Systematically assess and scale existing capability, including mature, documented pipelines with potentially reusable code, and NLP/LLM tools deployed in production.

Establish funding mechanisms for data engineering, including local infrastructure and recurrent funding for technical teams.

Establish a governance mechanism to maintain and evolve the specification as a living standard with clear revision processes.

Gap 2: Isolated data environments without scalable integration architecture to facilitate cross-environment analysis

Regional SDEs in England operate as isolated environments with no standardised cross-SDE analysis mechanism. Scottish Safe Havens lack automated federation. Welsh and Northern Irish infrastructure have no systematic connection to English or Scottish assets. Regional and organisational boundaries remain data boundaries.

No scalable integration architecture currently operates at population scale. Existing federated analytics investments have not transitioned to production services supporting routine researcher access across the full range of use cases. Researchers requiring data from multiple assets must navigate separate access processes, work in separate environments, and manually reconcile outputs.

Why is bridging this gap important for users?

Without integration, studies requiring data from multiple sources face restricted scope (reducing statistical power and generalisability), fragmented access with extended timelines, or relocation to jurisdictions where integrated access exists.

A pharmaceutical company conducting a post-authorisation safety study across all four nations (or multiple English SDEs) currently faces separate applications to multiple controllers. A clinical trial sponsor assessing feasibility for a rare disease trial cannot query patient counts against eligibility criteria across disparate trusts and nations simultaneously. An academic researcher investigating regional variation in treatment patterns cannot directly query comparable cohorts across regions.

Opportunity 2: A UK-wide data integration layer that enables cross-asset, cross-region, and cross-nation research at scale. Different architectural approaches exist, each with distinct trade-offs

Why this matters: Asset integration is the weakest layer in the UK landscape relative to HDRS ambition. No mechanism exists for cross-asset, cross-region, or cross-nation research at scale.

How it brings value: An integration layer provides the technical means by which HDRS could deliver its primary proposition: a single, predictable route to nationally representative health data for research.

What is the opportunity?

Establishing a data integration layer that allows approved researchers to discover and analyse data from multiple assets in combination, regardless of where those assets are held. This layer could work across organisational, regional, and national boundaries, enabling studies spanning multiple NHS trusts, English SDEs, devolved nations, and combinations of national assets, regional infrastructure, and specialist data sources.

An integration layer would connect existing assets rather than replace them. Data would remain where currently held, under existing governance arrangements, with local data controllers retaining responsibilities. The HDRS would provide connective infrastructure that allows distributed assets to participate in the same projects, with a single interface.

Three core functions could form part of this layer: (1) unified discovery, allowing researchers to understand what data exists across the network and assess study feasibility; (2) coordinated access, providing consistent routes to data held in distributed locations; and (3) combined analysis, enabling approved researchers to work with data from multiple assets within a single analytical environment.

Three broad architectural options exist: centralisation, federated analytics, and data fabric. Each has different characteristics, costs, and constraints that the HDRS would need to evaluate.

Centralisation physically consolidates data from multiple sources into a single repository. This offers simplicity for end-users and enables straightforward cross-asset analysis. However, centralisation requires data controllers to relinquish custody, which conflicts with legal responsibilities (particularly for GP data where practices are individual controllers), institutional interests, and public expectations. Previous UK attempts, including care data and aspects of GDPR, encountered governance objections and public opposition, and a new centralised infrastructure may carry a similar inherent risk of failure.

Federated analytics sends queries or algorithms to data rather than moving data to a central location. This preserves local custody and has achieved genuine capability in specific contexts: DataSHIELD for international epidemiological cohorts, OpenSAFELY for national primary care, and OMOP-based cohort discovery networks. However, traditional federation requires participating nodes to conform to rigid data schemas before integration is possible, and requires nodes to carry the burden of compute and service availability. Given UK heterogeneity, achieving universal conformance and reliability may delay meaningful integration, particularly where complex data types are involved. The different federated networks that currently exist are built using separate architectures, and would need substantial reconfiguration to participate in a new 'HDRS federation'. Strict 'eyes-off' approaches are prohibitive for use cases requiring data exploration, quality assessment, or regulatory-grade validation, and reduce statistical power when estimating effect sizes between groups if estimated in different nodes. However, use-cases where privacy is paramount (e.g., users prohibited from seeing row level data at all), aggregated queries (e.g., in discovery), and AI model training (federated learning), are a good fit for federation.

A data fabric uses a metadata layer (a layer that 'describes' data) which is broadly visible across a network, to provide unified discovery, access, and analysis across distributed assets. A user may write queries against disparate data assets through a single TRE, pulling transformed datasets into a secure environment for analysis. Data is maintained at source and not persisted in the TRE (though an analysis dataset can be recreated if needed). Unlike federation, a fabric does not require universal schema conformance before integration can begin – assets can be connected as their storage technology meets fabric specifications. However, a fabric could incorporate access to existing federated networks as part of its infrastructure. Unlike centralisation, it does not require custody transfer. However, there are no mature examples of data fabrics implemented cross-regionally in the UK. Fabric technologies are becoming more widespread, with SeRP evolving toward fabric

architecture, and cloud platforms increasingly adopted by NHS organisations that natively support this technology (NHSE SDE, South West SDE, Thames Valley and Surrey SDE, London SDE, Our Future Health, and others). These have compatibility with open source solutions, including open table formats and catalogues that are vendor-neutral, and are compatible with different processing engines.

The HDRS would need to determine which approach, or combination of approaches, best serves its objectives given constraints of time, cost, and end-user needs.

What value would this bring to the HDRS?

In the short term, connecting even a small number of assets would represent a step change in capability. Combined analysis across two or three English SDEs, leading hospitals, and four nations assets, in a way that delivers a reliable service, would provide population-level study capabilities where none previously existed.

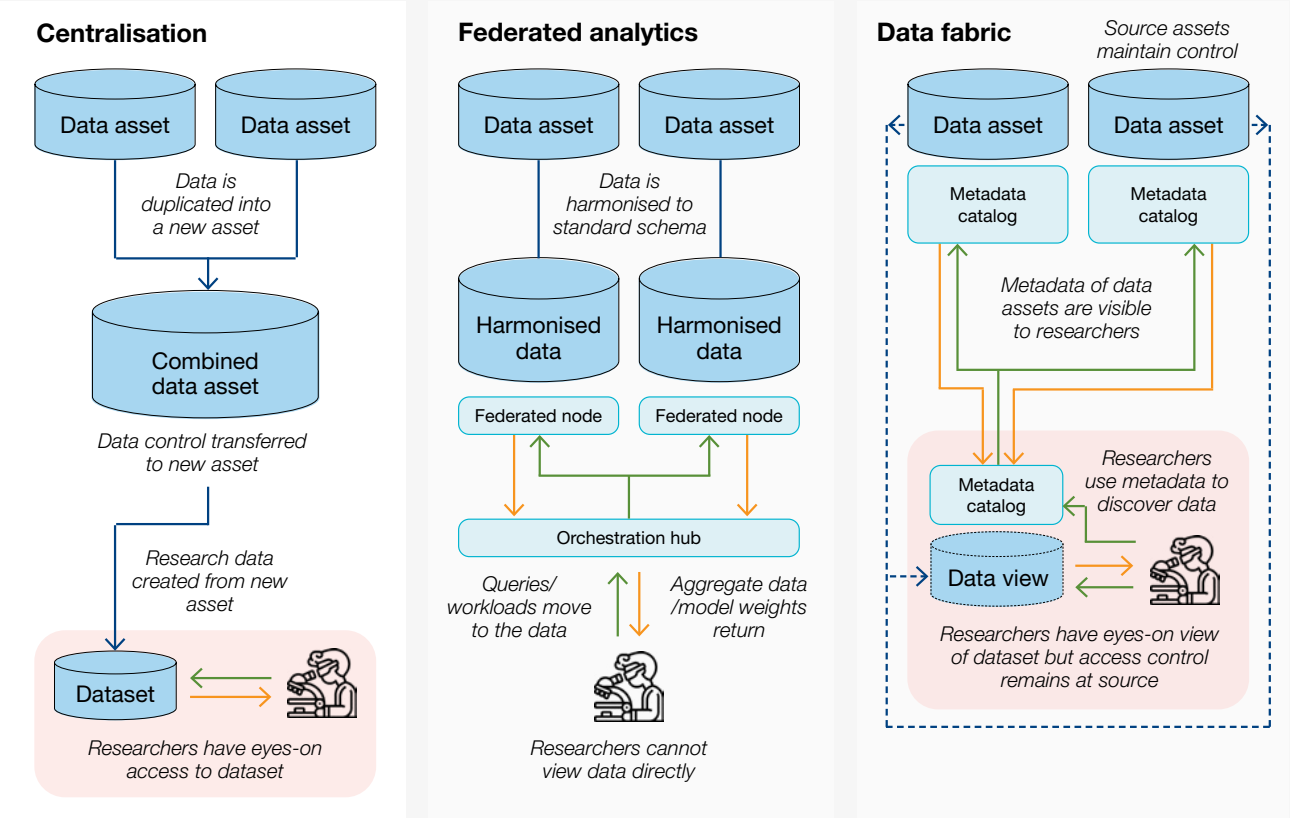
In the longer term, an integration layer becomes the core mechanism for HDRS value delivery. As more assets adopt compatible standards, the marginal cost of adding new assets falls. The HDRS could credibly offer commercial customers access to UK-wide representative data.

What is the reasoning behind this opportunity?

The UK is unlikely to have, in any realistic timeframe, a single national health data repository. Data is distributed across hundreds of organisations under different legal frameworks and technical environments.

The choice of architectural approach involves trade-offs. Centralisation offers simplicity but faces governance

Figure 7: Architectural options for data integration



Key

Data Queries Outputs TRE

Three approaches to asset integration, as described in Opportunity 2. Centralisation involves duplicating different assets into a new asset that persists, and independently serves researcher needs. Federated analytics involves keeping data entirely at source, and sending analysis code to data. This requires data to be precisely harmonised in all locations, and does not allow the researcher to see any data. A metadata-driven data fabric involves exposing metadata to researchers to allow discovery and querying of data. Datasets can be temporarily viewed and analysed in a TRE, but the original data always remains at source.

obstacles that have defeated previous attempts. Federation preserves local control and has demonstrated value for specific use cases, but requires schema conformance and excludes use cases requiring eyes-on exploration of data. Fabric approaches offer flexibility and allow incremental adoption, but require sustained technical leadership and do not eliminate underlying complexity.

A hybrid approach is another option: using federated mechanisms for discovery and feasibility queries where OMOP-standardised networks exist, while enabling combined analysis in TREs for use cases requiring eyes-on access. The HDRS would need to define how these mechanisms interact and what technical specifications assets must meet.

Building on open standards rather than proprietary platforms would reduce vendor dependency and enable assets to participate regardless of underlying infrastructure choices, though this does not eliminate technical complexity.

What are the risks of this approach?

Implementation requires sustained technical leadership to translate architectural patterns into concrete outputs that are directly tied to service delivery. This capability must be built or procured.

Assets would need to adopt compatible formats, implement required APIs, and maintain participating infrastructure. This requires investment and capability not all current assets possess. The integration layer would initially connect only those assets able to meet requirements.

Existing federated analytics investments represent substantial institutional commitment. Federation-first approaches need evaluation within a broader framework of HDRS end-user requirements.

An integration layer does not automatically solve data quality, governance, or access challenges. These require parallel investment.

What are the risks of prioritising other opportunities?

Without a data integration layer, HDRS options would be more limited: pursue centralisation (which has repeatedly failed), position existing federated analytics as the sole architecture (which excludes important use cases and requires universal conformance to exact data structures), or provide coordination without integration (discovery and governance services while leaving technical access fragmented).

Coordination without integration delivers some value but does not address the core gap. Researchers would still navigate separate environments, and cross-asset analysis would remain impractical. HDRS would function as a directory and governance framework rather than infrastructure enabling nationally representative research. The commercial proposition would weaken substantially.

If pursued, what are possible steps?

Define technical specifications for integration layer participation, including metadata standards, API specifications, and supported query patterns.

Evaluate architectural approaches against HDRS requirements, assessing trade-offs between centralisation, federation, and fabric patterns for different use cases.

Identify existing assets with emerging integration capability as candidates for early adoption and reference implementations.

Establish orchestration capability for unified interfaces to distributed assets, with initial scope focused on metadata aggregation, query routing, and enabling multiple datasets within analytical environments.

Integrate existing national assets and federated networks through appropriate pathways, whether direct technical integration or coordinated access.

Establish governance for evolving technical specifications as technology matures and new assets join.

Gap 3: Reliable, scalable pathways to data linkage

No mechanism exists to identify which data sources hold records for a given individual across the health system. Researchers cannot assess data availability without submitting speculative requests to multiple custodians.

Linking disease registries, biobanks, non-health records, and research cohorts to routine NHS records requires bespoke project-by-project arrangements. Governance is fragmented, with each linkage exercise typically requiring separate agreements, ethics approvals, and controller negotiations.

Where linkage occurs, quality is inconsistent and unmonitored. Different assets use different algorithms, thresholds, and approaches to uncertain matches. Researchers cannot quantify linkage error or compare results across studies using different methods.

Why is bridging this gap important for users?

Stakeholders and data custodians repeatedly reported that linkage is a substantial bottleneck in enabling reliable and timely research.

A precision medicine researcher linking genomic cohort data to NHS longitudinal records currently negotiates separately with each controller, uses ad hoc methods, and accepts unknown error rates. A pharmaceutical company conducting post-authorisation safety studies needs to link trial participants to long-term NHS outcomes, but no standard pathway exists with known quality and acceptable timelines. An academic researcher linking education records to health outcomes must develop bespoke linkage code, negotiate governance from first principles, and defend unvalidated methodology to reviewers. The project becomes a methods development exercise rather than substantive research.

Opportunity 3: Establishing national data linkage services with agreed methods, transparent quality metrics, and governance frameworks, to enable rapid, consistent linkage across distributed assets

Why this matters: Linkage is a critical failure point with no standard approach or quality metrics across the UK. Researchers cannot identify which sources hold records for

an individual, and linking research cohorts to NHS records requires bespoke arrangements adding months or years to projects.

How this brings value: Standardised linkage services become infrastructure that other HDRS capabilities depend upon, reducing project setup from months to weeks and enabling credible service level commitments to commercial customers.

What is the opportunity?

Establishing or accrediting linkage services that provide consistent, high-quality linkage of patient records across data sources. Such services would need to deliver three core functions: patient dataset resolution (identifying which data sources hold records for a given individual or cohort), deterministic linkage (matching on NHS number or equivalent identifiers), and probabilistic linkage (matching on demographic variables for sources lacking reliable identifiers).

Several options for delivery models could achieve this. A single national linkage service would provide consistency but risks creating bottlenecks and duplicating existing regional capability. A network of accredited providers meeting common standards could leverage existing investments (NHS England Master Person Service, SAIL, Scottish Safe Havens, Honest Broker Service) while ensuring interoperability. A hybrid model could establish central coordination with distributed delivery. The choice involves trade-offs between consistency, capacity, and respect for existing capability.

Regardless of model, effective linkage services would benefit from pre-agreed governance frameworks with standard data sharing agreements and clear legal bases, and mechanisms to apply National Data Opt-Out preferences and other consent restrictions. Linkage services that regularly published algorithms and transparent quality metrics (match rates, false positive/negative estimates, coverage statistics) would improve auditability and accountability. There is potential for linkage services to be accessible through an integration layer.

What value would this bring to the HDRS?

Linkage services could function as infrastructure underpinning other HDRS capabilities. A data integration layer requires consistent patient resolution for cross-asset queries. Data asset value improvement depends on linkage connecting hospital prescribing and laboratory data to primary care.

In the short term, defined quality metrics and pre-agreed governance has potential to reduce project setup from months to weeks.

In the longer term, patient dataset resolution would enable efficient study design, with researchers identifying relevant assets before committing to applications. Published quality metrics would enable researchers to account for linkage error and allow systematic algorithm improvement.

What is the reasoning behind this opportunity?

The NHS number provides a strong foundation for deterministic linkage, with high coverage across primary and secondary care. Scotland's CHI and Northern Ireland's HSC number provide equivalent capability. This is a UK strength relative to systems lacking universal identifiers.

However, deterministic linkage alone is insufficient. Research cohorts established before routine NHS number capture, disease registries with historical data, and non-health sources (education, social care) require probabilistic linkage on demographic variables.

Probabilistic linkage is technically mature but operationally fragmented. SAIL has demonstrated robust probabilistic methods at population scale. NHS England operates the Master Person Service. Scottish Safe Havens have developed capabilities, though cross-regional projects rely on manual coordination.

These represent genuine but inconsistent capabilities using different algorithms, different types of data, quality thresholds, and governance arrangements. Linkage is generally built on specific data assets, and is not available as a uniform service across the UK. Many data assets are currently unlinkable.

What are the risks of this approach?

Linkage across sources raises privacy concerns requiring careful governance. In PPIE sessions we found that public perception of linkage as surveillance may create opposition, requiring clear communication of research purpose and safeguards.

Deeper linkage increases re-identification risk, particularly for small populations or rare characteristic combinations. Services would need safeguards with potential restrictions on particularly sensitive combinations.

Linkage services require transient access to identifiable data, concentrating security risk and requiring stringent information security standards on the organisation performing the linkage.

Pre-agreed governance frameworks require negotiation across controllers, regulators, and devolved administrations. Edge cases would still require bespoke arrangements.

Cross-nation linkage introduces complexity. Unique identifiers are not automatically interoperable with no UK-wide crosswalk. Northern Ireland lacks secondary legislation enabling identifiable data release without consent. Some cross-nation scenarios may remain impractical near-term.

What are the risks of prioritising other opportunities?

Without standardised services, every cross-source project solves linkage independently, consuming researcher time, introducing inconsistent methods, and creating delays reducing UK competitiveness.

Linkage quality remains unknown and variable. Researchers cannot account for or quantify linkage error. Studies using different methods cannot be meaningfully compared.

High-value sources remain isolated. Biobanks cannot be reliably linked to NHS records. Research cohorts cannot be followed through routine data. Integration underpinning precision medicine and life-course epidemiology remains impractical at scale.

The alternative is treating linkage as out of scope, leaving it to individual projects and controllers. This preserves status quo conditions identified as a structural barrier.

If pursued, what are possible steps?

Conduct comprehensive assessment of existing linkage capabilities across the UK, documenting methods, algorithms, quality metrics, governance, capacity, and turnaround. This includes NHS England services, SAIL's anonymous linking field approach, Scottish Safe Haven functions, Northern Ireland's Honest Broker Service, regional SDEs, the UK LLC, and open-source tools such as the Ministry of Justice 'Splink' library.

Define standards for linkage service accreditation covering algorithm validation, quality reporting, information security, governance, and service levels, accommodating both deterministic and probabilistic methods.

Establish or extend governance frameworks providing pre-agreed legal bases for common scenarios: research cohort linkage to NHS data, cross-nation linkage, and non-health source linkage.

Develop patient dataset resolution capability enabling researchers to understand data availability before committing to specific requests.

Implement continuous quality monitoring with published metrics at defined intervals and mechanisms for algorithm improvement.

Gap 4: Variability in TRE capability

TRE quality and capability varies substantially. Some environments offer comprehensive software libraries, responsive support, and adequate compute. Others provide minimal tooling, limited support, and infrastructure unsuited to large-scale workloads.

Current accreditation frameworks assess security requirements but do not systematically evaluate researcher needs. Compute co-location with data remains unresolved – high-performance computing facilities do not have processes in place for persisting sensitive data or managing the research project lifecycle.

Why is bridging this gap important for users?

Stakeholders report that TRE experiences and capabilities can restrict ability to conduct research. This may limit what can be done on otherwise valuable data assets, or limit what research can be performed in the UK in general.

A pharmaceutical company with an approved data access agreement discovers the TRE lacks specified statistical software, requires weeks to install packages, and runs analyses slowly. An AI developer requiring GPU compute to train models on imaging data finds the hosting TRE has no GPU capability. An academic researcher accustomed to one regional SDE receives access to data held in another – different software versions, file systems, and interfaces mean code developed in the first environment does not run in the second.

<u>Opportunity 4: TRE accreditation standards addressing usability alongside security, with options for how HDRS guarantees consistent researcher experience across the network</u>

Why this matters: TRE quality varies substantially – some offer comprehensive tooling and compute, others impede research through poor usability. Researchers cannot predict what environment they will encounter.

How this brings value: Consistent standards enable credible service level commitments. Addressing compute co-location unlocks AI and computationally intensive workloads currently impractical at scale.

What is the opportunity?

TREs are the accepted model for secure health data research access in the UK. However, current accreditation focuses on information security, not researcher experience. A TRE can be accredited as secure while offering poor usability. The opportunity is to establish standards that address both dimensions, creating accountability for the environments through which users access HDRS data.

Several design options require consideration:

- **Scope of standards:** Accreditation could extend beyond security (ISO 27001, SATRE, UKSA) to include usability criteria: software availability, compute capability, support responsiveness, and service levels. Alternatively, HDRS could maintain security-only accreditation and address usability through guidance or contractual requirements with individual TREs.
- **Central TRE:** HDRS could operate its own TRE as a guaranteed access point, ensuring the service controls infrastructure required to deliver commitments. Alternatively, HDRS could function purely as a coordination layer, relying entirely on existing regional and specialist TREs. A hybrid model would maintain a central TRE alongside accredited regional alternatives.
- **Compute co-location:** Options include integrating supercomputer environments (Edinburgh Parallel Computing Centre, Dawn, Isambard-AI) as accredited TREs for health data, requiring accredited TREs to provide or connect to HPC/GPU resources, or treating compute as a separate infrastructure stream outside TRE accreditation.
- **Network composition:** Accreditation could be limited to public-sector TREs, or extended to commercially owned environments where researchers may wish to analyse commercial and NHS data together.

What value would this bring to the HDRS?

Usability standards would create accountability for researcher experience. Users could be confident that

accredited environments meet baseline expectations, reducing project risk and improving time-to-analysis.

A central TRE (if pursued) could enable HDRS to make credible service level commitments, respond to user feedback, and provide access for users without existing regional relationships. It would ensure HDRS is not wholly dependent on infrastructure it does not control.

Addressing compute within the TRE framework reflects the user perspective: researchers need environments where they can analyse data, and compute capability determines fitness for purpose.

What is the reasoning behind this opportunity?

Stakeholder engagement identified consistent frustration with TRE usability. Researchers reported environments where basic software was unavailable or required lengthy approval. Compute constraints were cited as barriers. The variation itself was problematic – researchers could not plan projects with confidence.

The case for a central TRE rests on the principle that a service should control infrastructure required to deliver or de-risk its commitments. Users seeking UK-wide data access currently have no single environment designed for that purpose.

The case for maintaining regional accreditation recognises existing investments and legitimate preferences for local infrastructure. Forcing all access through a central environment would create bottleneck risk, duplicate capability unnecessarily, and generate resistance from regions that have invested in their own infrastructure.

What are the risks of this approach?

Usability standards introduce subjectivity that security standards avoid. Disagreement may arise about what software should be pre-approved, what compute is adequate, or what response times are acceptable.

Existing TREs failing to meet standards may be excluded from the network, creating tension with data controllers who have invested in local infrastructure. Accreditation must be positioned as a pathway to participation, not a barrier.

A central TRE could become a bottleneck or single point of failure. Capacity planning must anticipate growth. Regional TREs provide resilience if the central option is pursued.

Regional SDEs may view a central TRE as competition for users, funding, and relevance. Clear frameworks and communication about complementary relationships would be essential.

What are the risks of prioritising other opportunities?

Without usability standards, TREs remain a source of unpredictable researcher experience. Users with alternatives will place projects elsewhere. HDRS cannot make credible service level commitments if it cannot guarantee what users will experience.

Without a guaranteed access point under HDRS control, the service depends entirely on infrastructure operated by others with different priorities. If regional TREs cannot meet demand or lack required capabilities, HDRS has no fallback.

Without addressing compute co-location, AI and computationally intensive research remains impractical at scale. The UK cannot compete for machine learning workloads requiring GPU access to large datasets.

The alternative is accepting existing TRE variation and focusing HDRS on data integration alone. This preserves local autonomy but perpetuates inconsistency and makes HDRS dependent on infrastructure it does not control.

If pursued, what are possible steps?

Define minimum security standards and architecture requirements, drawing on existing frameworks (ISO 27001, SATRE, UKSA, DHSC).

Define usability criteria covering software availability, compute capability, support responsiveness, and service levels, developed with input from researchers across user segments.

Determine whether HDRS should operate an environment for deploying TREs and, if so, evaluate options for infrastructure (e.g., building on NHS National SDE, NHS Supercomputer Network, SeRP, or new procurement), operating model (in-house or contracted) and customisability (one per use case, approved registry of docker images, bespoke configuration each time).

Establish technical integration specifications between TREs and any data integration layer, including APIs for discovery, request, delivery, authentication, and audit logging.

Assess existing TREs against proposed standards, informing both accreditation pathways and decisions about central TRE capability scope.

Determine policy on commercially owned TRE participation in the network.

Establish processes for ongoing standard maintenance as technology and user requirements evolve.

Gap 5: Access processes as a key barrier to delivering at pace

Access processes are repeatedly cited as a substantial barrier to research. The current system operates through sequential approvals that are unpredictable. Studies can pass ethics review and still be rejected later (e.g., due to varying requirements between organisations), making it possible to fund undeliverable research. Identical governance work is repeated across organisations. Researchers report each data controller requires bespoke documentation even when underlying requirements are substantively identical.

Research project steps are managed through separate systems with no integration. Progress is tracked through emails and spreadsheets. Approvals in one system are not visible in another. Researchers cannot predict access timelines, and data controllers cannot demonstrate performance or identify bottlenecks.

A Safe People Registry exists but is under-invested and not universally recognised. Researchers who have completed training must re-establish credentials with each new data controller.

Why is bridging this gap important for users?

Commercial users report that inability to guarantee timelines makes UK data access a risk, and projects are placed elsewhere because access lacks predictability.

Academic researchers applying to multiple controllers face separate documentation in different formats, credentials not recognised across controllers, and institutional agreements requiring renegotiation. Clinical trial sponsors linking participants to NHS outcomes face separate applications for each data source, each with different forms, fees, and timelines. Patients and the public cannot easily see what projects have been approved, what data was accessed, or what outputs resulted.

Opportunity 5: Accelerating access through solutions like a unified digital governance layer, incorporating portable researcher credentials, reusable templates, and transparent project lifecycle tracking

Why this matters: Governance processes are fragmented, manual, and opaque. Identical work is repeated across organisations, timelines are unpredictable, and researcher credentials are not portable.

How this brings value: Standardised processes and visible performance metrics create accountability, reduce duplicative burden, and enable HDRS to credibly commit to service levels.

What is the opportunity?

A unified digital governance layer could standardise research access from initial enquiry through to data use reporting, providing a single environment where researchers apply, data controllers process requests, and all parties track project status.

Three components are options for being part of this layer. First, portable researcher credentials where individuals demonstrate training, identity verification, and institutional affiliation once, with recognition across participating data controllers. Second, reusable machine-readable templates for data sharing agreements, access agreements, and commercial contracting that can be configured without bespoke legal drafting. Third, project lifecycle tracking that provides visibility to researchers on application status, to data controllers on pipeline and performance, and to the public on data use.

Additional options exist for how this could be delivered. One approach builds on existing tools developed by HDR UK and others, including the Five Safes application framework, template agreements, the Safe People Registry, and the Data Use Register Standard. This leverages prior investment but requires assessment of whether these tools are fit for purpose at scale. An alternative approach develops new infrastructure designed for HDRS requirements, avoiding constraints of legacy systems. A hybrid approach might adopt proven components while replacing those that do not meet requirements.

What value would this bring to the HDRS?

In the short term, unified tracking can provide visibility that creates accountability. Researchers see application status. Data controllers identify bottlenecks. HDRS reports aggregate metrics on access times and approval rates. Portable credentials reduce duplicative verification. Reusable templates accelerate contracting without negotiation from first principles.

In the longer term, a governance layer could become a mechanism through which HDRS can credibly commit to service levels. If timelines and performance are visible and auditable, HDRS can identify where the system fails expectations and direct support accordingly. The system supports both public transparency and researcher transparency.

What is the reasoning behind this opportunity?

Stakeholder interviews consistently identified governance unpredictability as a barrier to UK competitiveness. Current processes are manual, and confined to numerous versions of forms, documents, and spreadsheets.

Transparency of performance is preferable to attempts to directly enforce an approach as this reflects the reality of distributed data controllership. Controllers have legal responsibilities that cannot be overridden by HDRS mandates. However, transparency creates accountability: if one controller processes applications in weeks while another takes months, this becomes visible to researchers, funders, and policymakers.

What are the risks of this approach?

Standardised templates may not accommodate legitimate variation. Some projects genuinely require bespoke arrangements due to novel data types or complex structures. The system must provide pathways for non-standard requests.

Centralised tracking of credentials and applications creates a sensitive data asset requiring robust security and agreement over who should hold that data.

Adoption requires data controllers to change established processes. Controllers with functioning local arrangements may resist migration, particularly if transition imposes short-term costs. Benefits to controllers must be demonstrated alongside benefits to researchers.

Visible performance differences between controllers may create political tension. HDRS must be prepared to support improvement rather than simply expose problems.

What are the risks of prioritising other opportunities?

Without a unified governance layer, access remains opaque and unpredictable. Commercial users place time-sensitive work elsewhere. Researcher credentials remain non-portable.

Agreement templates exist but are not universally adopted. Performance remains invisible, preventing identification of bottlenecks or credible service level commitments.

The alternative is focusing HDRS investment on data and technology while leaving governance to individual controllers. This avoids complexity of changing established processes but accepts that access unpredictability will continue to constrain the value that improved data and infrastructure can deliver.

If pursued, what are possible steps?

Conduct requirements analysis across user segments to identify minimum viable functionality, distinguishing essential features from later enhancements and weighting organisations which handle large volumes of research over less mature services.

Assess existing tools including HDR UK templates, the Safe People Registry, and data controller systems. Identify what can be adopted, what requires modification, and where new development is needed.

Define the data model for portable credentials, specifying what constitutes validity, how credentials are verified and maintained, and how cross-controller recognition is operationalised.

Develop reusable, machine-readable agreement templates covering common scenarios, with legal review confirming templates meet requirements across all four nations.

Specify project lifecycle tracking including defined stages, status transitions, and reporting capabilities for both researchers and system-wide performance.

Pilot with a subset of data controllers holding key assets and willing to adopt new processes, gathering feedback before broader rollout. Pilot selection should include controllers from multiple nations.

System opportunities

The technological opportunities described in the previous section address gaps in data, infrastructure, and access. However, the gap analysis and stakeholder engagement identified constraints that sit outside the technical domain, but which are key enabling conditions that determine whether technological investments can deliver their intended value.

This section highlights the most pervasive non-technical gaps in commercial, funding, governance, and public and patient trust that emerged from the review, and presents opportunities to bridge these gaps, drawn from existing practice and evidence.

System Gap 1: Funding and service models that enable sustainable and scalable research infrastructure

The current UK health data ecosystem produces innovation but struggles to sustain it or translate it into meaningful impact. Academic funding cycles prioritise novelty and publications, over service capabilities, maintenance, stability, and end-user support. NHS funding cycles prioritise strategic infrastructure, but rarely specify end-user and impact focused delivery objectives (i.e., an infrastructure is not a success by existing, but must generate measurable value).

When projects end, teams disperse, code accumulates technical debt, and institutional knowledge disappears. Technical assets frequently depend on individual staff members, creating single points of failure. Previous substantial investments have often built useful prototypes, but without dedicated engineering pathways and product focused design, these do not scale into reliable national infrastructure.

<u>Supporting Opportunity 1: Transition the ecosystem from grant-funded projects to a professional service model</u>

Why this matters: Critical research infrastructure depends on short-term academic or NHS funding without clear service delivery objectives, resulting in fragility, lost institutional knowledge, and lack of clear impact when funding periods end.

How this brings value: A professional service model creates stable, production-grade infrastructure with predictable performance, enabling HDRS to compete for commercial contracts and make credible service commitments.

The opportunity is to separate core data infrastructure – ability to deliver to end-user requirements, reliability, and on-going support – from activities that drive novelty or siloed development. Core assets could be treated as long-term products measured for service success, rather than short-term research and technical outputs. This may require recognising data engineering and software/platform engineering as professional disciplines within the health data ecosystem, with career pathways, appropriate remuneration, and institutional recognition.

The option for a professionalised model could enable HDRS to set and enforce service level agreements, fund maintenance alongside development, implement version control and change management, and build teams that persist beyond funding cycles. Some national assets such as CPRD and SAIL have demonstrated success through this approach: dedicated engineering teams treating data provision as a long-term service rather than a temporary research output. There is also the option to capture and protect intellectual property developed through public investment, ensuring foreground IP (code, algorithms, validated pipelines) remains available as national assets rather than being lost or locked into proprietary platforms.

The risks of this approach include tension with academic institutions whose incentives differ, and the challenge of building engineering capacity in a competitive labour market. The risks of not pursuing it include the inability to offer the predictable service that commercial users require, continued fragility, and repeated loss of capability when grant cycles end.

System Gap 2: Lack of incentives to implement consistent standards and technology for research readiness for data controllers and custodians

Data controllers and custodians of existing assets invest substantial effort in making data research-ready: cleaning, transforming, documenting, and maintaining datasets while navigating governance requirements. Many of these functions cannot be pushed downstream, so some costs and risks will inevitably remain local. Without clear returns, controllers rationally prioritise local operational needs over national research infrastructure.

Supporting Opportunity 2: Establishing a value-return framework

Why this matters: Local data controllers bear costs and risks of preparing research-ready data without reliable mechanisms to secure fair value in return, creating misaligned incentives.

How this brings value: A value-return framework creates sustainable revenue streams for data controllers, incentivising participation in national infrastructure and enabling predictable pricing for users.

The opportunity is to incentivise building towards HDRS infrastructure patterns, by implementing mechanisms that return tangible value to participating data controllers. Options include direct revenue sharing from commercial access fees, on-going licensing of data assets, investment in local data engineering capacity, access to tools and validated code developed through the network, or operational benefits such as benchmarking data and quality insights. A framework could build on the HDR UK Gateway to create standardised service listings, transparent pricing tiers, and master agreements that reduce negotiation burden.

For commercial users, a centrally managed framework could replace opaque, lengthy negotiations with predictable timelines and published pricing. Industry stakeholders indicated willingness to pay premium rates for guaranteed access within defined timeframes. For academic users, tiered pricing supported by commercial revenues could maintain affordable access while ensuring the system is financially sustainable. For controllers, predictable revenue streams would support the engineering teams required to maintain research-ready assets, reducing dependence on grant funding.

Central transaction tracking would enable performance monitoring across the network, identifying where bottlenecks occur and where investment is needed. This visibility supports continuous improvement and creates accountability that current fragmented arrangements lack.

The risks include complexity in defining fair value allocation across heterogeneous controllers, and potential resistance from organisations with established commercial arrangements. The risks of not pursuing it include continued misalignment between those who bear costs and those who capture value, undermining participation in national infrastructure.

System Gap 3: Sustained public trust

Health data research operates with patient and public permission. Surveys consistently show broad support for using NHS data for research that benefits patients and the health system, but this support is conditional.

Previous initiatives, for example care.data and GPDPR, have collapsed where the public perceives that commercial interests were prioritised over patient benefit and that data use lacked transparency and accountability. The National Data Opt-Out exists because trust was damaged. Approximately 5.4% of the population has opted out, with rates spiking during trust crises.

Supporting Opportunity 3: Secure social licence through demonstrable trustworthiness and transparency

Why this matters: Public trust is the foundation on which health data research and infrastructure development depends. Previous initiatives have failed when this trust was lost.

How this brings value: Embedded transparency and public engagement create durable social licence, reducing risk of backlash that could halt HDRS operations entirely.

The opportunity is to build transparency and public accountability into HDRS operations from the outset, rather than treating them as afterthoughts. Options include public-facing tools that explain in accessible language who accesses data, for what purposes, and what outcomes result. It could include standing governance bodies with genuine patient and public representation that address contested questions about commercial access, equity, and data use in sensitive areas. It could include mechanisms that demonstrate value returning to the NHS and patients – not just economic growth figures, but tangible improvements in care. There is opportunity for activity reporting to be automated, for both customers of the service and the public.

Stakeholder engagement and patient workshops identified several priorities: clear communication about safeguards and red lines, consistency in how data is governed across the system, and inclusion of diverse voices in decision-making. International examples demonstrate that radical transparency is feasible – Estonia's X-Road publishes access logs showing citizens exactly who has viewed their data.

The risks of this approach include the resource requirements of genuine engagement (as opposed to tokenistic consultation), and the possibility that transparent performance data reveals inefficient processes that create short-term reputational damage. The risks of not pursuing it include those that are existential: if public trust collapses, the HDRS proposition may become undeliverable. No amount of technical excellence can compensate for lost social licence.

System Gap 4: Incompatible local systems and vendor lock-in

NHS organisations make independent procurement decisions for clinical systems, data platforms, and analytical tools. These decisions are typically optimised for local requirements without regard to national interoperability. The result is a landscape of incompatible systems where data cannot flow between organisations, where each integration requires bespoke technical work, and where vendors can

lock data behind proprietary interfaces. Long-term contracts for legacy systems further constrain modernisation.

Supporting Opportunity 4: Align technology procurement with national technical standards

Why this matters: Fragmented local procurement creates technical silos, vendor lock-in, and incompatible systems that obstruct national integration.

How this brings value: Coordinated procurement enables interoperability, reduces costs through aggregated purchasing power, and prevents data being locked behind proprietary barriers.

The opportunity is to establish procurement frameworks that require alignment with national technical standards as a condition of purchase. This does not mean mandating specific products – local teams would retain choice among compliant options – but ensuring that whatever is purchased can participate in national infrastructure. Standards could address data formats, API specifications, interoperability requirements, and intellectual property terms ensuring that code and algorithms developed using NHS data remain accessible rather than becoming vendor property.

For local organisations, coordinated procurement could reduce negotiation burden and provide access to nationally negotiated terms. For HDRS, it could create a technical environment where integration is possible without bespoke engineering for each connection. Aggregated purchasing power has potential to reduce costs for cloud computing, storage, and common tools, making large-scale research more affordable.

The Procurement Act 2023 introduces mechanisms such as the Competitive Flexible Procedure that could support modular partnerships and innovation rather than rigid long-term contracts. The NHS Value Sharing Framework provides precedent for ensuring public benefit from commercial arrangements.

The risks include resistance from vendors with established market positions, and from local organisations with existing arrangements they wish to protect. Transition costs may be substantial where legacy systems are deeply embedded. The risks of not pursuing it include perpetuating technical fragmentation that makes national integration progressively harder as incompatible systems multiply.

System Gap 5: Systems that enable initiation of clinical trials at an acceptable pace

Government targets call for 150-day trial setup by March 2026 and quadrupled participant recruitment by 2029. Current performance is moving in the wrong direction: industry-sponsored trial enrolment reached a seven-year low in 2024/25[27]. Stakeholders consistently describe the UK as "an unreliable and unpredictable partner," with one major pharmaceutical company reporting the UK as second slowest of 18 European countries for trial setup.

The underlying problem is that systems for patient identification, eligibility assessment, and recruitment only address parts of the requirement. Feasibility tools that query administrative data do not expose clinical depth – staging, biomarkers, prior treatment lines, exclusion criteria. Precision medicine trials stratifying by molecular markers cannot be served by infrastructure designed for population-level case finding. For oncology trials up to 90% of patients identified through current methods do not meet detailed eligibility criteria at pre-screening.

Supporting Opportunity 5: Establish a national clinical trial recruitment service that is integrated with separate data assets and local teams

Why this matters: UK clinical trial performance has declined significantly, with the country falling from 4th to 8th globally for Phase III industry trials, with slow and unpredictable recruitment being a key factor.

How this brings value: Integrated recruitment infrastructure addresses a high-value commercial market while demonstrating HDRS capability to deliver measurable improvements in research delivery.

Clinical trials represent both a commercial opportunity and a test of whether UK health data can translate into research delivery.

The opportunity is to build a recruitment service capability through HDRS that addresses these gaps. This would require integrating data assets that provide clinical depth (hospital records, cancer registries, laboratory results) with existing pathways for population-level trials on simpler criteria (for example – vaccine trials). Critically, it would require expansion of pathways that engage clinical teams in recruitment rather than relying solely on mass patient communication. CPRD's model demonstrates that clinician-mediated data-driven recruitment achieves higher conversion rates, but this capability does not exist for hospital-based trials.

An HDRS recruitment service could build on the same integration mechanisms that would be used to access disparate assets for analysis. The service would function as a 'research concierge' managing the end-to-end journey from feasibility query to connecting sponsors to local clinicians.

For trial sponsors, integrated infrastructure would provide feasibility assessment against detailed eligibility criteria (not just administrative proxies), predictable recruitment timelines backed by service level commitments, and reduced operational burden through standardised site engagement. For patients, integration with disparate systems could extend trial opportunities and make access more equitable. For HDRS, clinical trials offer a measurable capability demonstration with clear commercial value.

The risks include governance complexity in establishing HDRS as a data controller, where necessary, for recruitment activities, and the challenge of integrating systems and existing services. The risks of not pursuing it include continued decline in UK competitiveness for clinical trials, with sponsors placing studies in jurisdictions offering faster, more predictable recruitment.

Pilot initiatives

Pilots can serve three purposes. First, they validate that technology components work as intended when combined to serve real research needs. Second, they demonstrate value to users, funders, and the public within realistic timeframes, creating evidence that HDRS investment generates returns. Third, they surface implementation challenges that cannot be anticipated through design alone, enabling course correction before national rollout.

The pilots described in this section are illustrative examples only. They have been produced based on stakeholder input from our engagement, designed to demonstrate how the change principles described in this paper may be applied to develop real-world assets. They are not intended to represent the HDRS team's emergent views or delivery priorities – pilot selection should be informed by further detailed feasibility assessment, stakeholder engagement, commercial interest, and available resources. The examples presented here demonstrate a range of capabilities that could be enabled through technology available today, and are not the only pilots that could deliver value. Where we believe there are obvious case studies that could support these ideas we have outlined them, although these are by no means exhaustive.

A principle underlying this approach is that infrastructure should follow delivery, not precede it. The most effective route to building national capability may be through high-value projects that require infrastructure to be developed as a dependency. Projects with clear user demand, measurable outcomes, and commercial or policy urgency create forcing functions that drive infrastructure development toward genuine requirements. The pilots below are framed accordingly: each addresses a concrete use case while requiring technical capabilities that would become reusable national assets.

Pilot 1

Real-world evidence asset linking GP, hospital, and prescribing data

This pilot would establish a linked dataset combining primary care records, secondary care data, and hospital prescribing information for a population representing at least 30% of UK residents. The asset would enable population scale studies that track medication exposure across care settings, from GP prescription through hospital administration, linked to clinical outcomes.

This addresses a specific gap: hospital prescribing data, where high-risk and specialist medications are initiated, is largely unavailable at population scale. Real-world evidence studies therefore proceed with incomplete exposure data or are placed in jurisdictions where linked data exists. High-cost drugs data is already collected at each ICB, and as other reviews have stated this should be made available nationally at the earliest opportunity. This would also focus on hospital prescribing data outside of this collection.

The technical challenge centres on hospital prescribing extraction. Primary care prescribing is available through established assets, but hospital medications data is found in trust pharmacy systems. Between 2020 and 2023 this data was collected weekly in England through an integration with CareFlow Medicines Management covering 15% of trusts. In Scotland the same data collection has continued. This pilot would test whether source data engineering investments (Opportunity 1) can unlock this data type widely and uniformly, whether a combined asset can be made available through an integration layer (Opportunity 2), and whether national linkage services (Opportunity 3) can reliably join records across settings. The 30% coverage threshold represents a minimum viable population for adequately powered real-world evidence studies with external validity but is being reached by SDEs' and four nations' infrastructure with primary care data integration.

Success would be measured by the volume of pharmaceutical and regulator-led projects using the linked asset, with project timelines competitive against international alternatives.

User stories served: pharma customer, regulator, academic researcher

HDRS capabilities demonstrated: access to comprehensive health records, linking data for greater impact

Where to start: This pilot requires a standard definition for the data that is in scope. Starting points for this could include SCoMeD, the recently released dataset linking primary and secondary care prescriptions in Scotland. Prioritising secondary care integrations where pre-existing primary care information governance exists for geographically complete populations could provide a faster route to 30% population coverage – for example, through English SDEs with GP data flows, and SAIL in Wales. Integrations may be accelerated in geographies where single vendor EHR systems are in use in secondary care – for example, Epic in Northern Ireland.

Pilot 2

UK-scale biomarker-enriched pan-cancer cohort for clinical trials

This pilot would create a UK-wide oncology research data asset, spanning multiple tertiary cancer centres across the

four nations, with sufficient clinical depth to support trial feasibility assessment, site selection, and patient identification. Data would include staging information, lines of therapy, molecular markers, and treatment response extracted from clinical documents using NLP. Linkage to mortality outcomes would also support real-world evidence research.

Trial sponsors report that the UK cannot provide systematic feasibility data for biomarker-stratified trials, restricting placement of precision oncology studies. The clinical detail required for eligibility assessment currently exists in oncology records but is not extractable at scale.

Such a pilot would test how source data engineering (Opportunity 1) and an integration layer (Opportunity 2) can combine to deliver a functioning clinical trials data asset. This must be coupled with service processes for liaising with trials centres on setup and recruitment.

Success would be measured by commercial trial sponsors completing feasibility studies within defined service levels, with demonstrable improvements in time-to-recruit and reduction in screen failure rates.

User stories served: clinical trial sponsor, pharma customer, regulator, precision medicine researcher

HDRS capabilities demonstrated: opening up advanced diagnostics data, faster clinical trials

Where to start: This pilot relies on direct integrations with source systems at hospitals that have already begun to curate their cancer data and can access identifiable data in free text to enrich their records. DATA-CAN is an existing network that promotes collaborative activities across four nations, and could support UK-wide activity. Work from previous cancer data curation and NLP technology activities funded through the SDE network could be reused. A 'quick win' could be to make National Disease Registration Service data available for national clinical trials recruitment, but this data would still need to be enriched with 'deeper' data at site level. Any implementation would require either a new clinical trials service wrapper that focuses on specialist centres, or be able to adapt clinical trials' recruitment functions held by existing organisations.

Pilot 3

UK population epidemiology asset with cross-sector and cross-border linkage

This pilot would link health data to non-clinical determinants (such as social care, education, housing, environmental exposures) for at least 30% of the UK population, to enable new population research capabilities. It would also establish technical and governance pathways for studies requiring data from multiple nations. Achieving a flagship UK-wide study could serve to illustrate UK potential for conducting this type of work at population scale.

Such a pilot would test national linkage services (Opportunity 3) under demanding conditions: probabilistic matching where NHS number is absent, linkage across devolved administrations with different identifier systems, and governance frameworks spanning multiple data controllers.

Success would require demonstrating that cross-border, cross-sector studies can be established within predictable timescales, with linkage quality metrics supporting valid epidemiological inference. The four nations element is essential: a pilot achieving coverage only in jurisdictions with mature infrastructure would not test the difficult problems.

User stories served: academic researcher

HDRS capabilities demonstrated: access to comprehensive health records, linking data for greater impact

Where to start: This pilot would need to define a technical standard approach for a cross-sector linkage model, and define a UK-wide governance framework for cross-border linkage. Core datasets that are both accessible and add value to population wide studies, such as hospital activity and social care, could be targeted in a starting phase. An initial approach could be to explore whether different existing linkage services could be funded to maintain the same technical approaches, operating under service level agreements, to derisk delivery and outcomes. The Welsh SAIL team and the UK Longitudinal Linkage Collaboration already operate at population scale and could be a starting point for exploration.

Pilot 4

Sovereign AI foundation model including linked digital pathology

This pilot would position HDRS as the owner of a national AI initiative, providing linked imaging and digital pathology data to support development of a foundation model. A general foundation model can support clinical diagnostics, risk prediction, as well as drug discovery and development, and represent new intellectual property.

The pilot would need to establish pathways for large-scale imaging data to be analysed on national supercomputing infrastructure, testing whether compute co-location requirements (Opportunity 4) can be met for AI workloads that cannot practically operate within current TRE configurations. Disease focus should be selected based on data availability, clinical priority, and alignment with national strategy.

This is technically ambitious, requiring integration of imaging data pipelines, supercomputing access, and governance frameworks for AI model development. It tests capabilities that do not currently exist at scale: no large NHS imaging datasets have run on the UK supercomputer network, and the pathway from research model to clinical deployment remains undefined.

Success would be measured by model development proceeding to defined milestones, with the data access and compute pathway functioning as repeatable infrastructure rather than bespoke arrangement.

User stories served: medtech/biotech developing AI-based tools, precision medicine researcher.

HDRS capabilities demonstrated: opening up advanced diagnostics data, linking data for greater impact.

Where to start: A necessary starting point is working with the UK AI Research Resource network to agree on standard technical and governance frameworks to run sensitive health data studies on national supercomputing infrastructure. Multimodal data for research could be made available at site-level, including in large teaching hospitals across four nations, and through existing national networks such as NPIC in England (digital pathology) and iCAIRD in Glasgow (radiology and pathology). New infrastructures for deployment, such as pilot and national roll outs of prostate and breast cancer pathology screening AI in Wales, and the new UK AIR-SP, could provide an opportunity to define standardised pathways to AI deployment and evaluation.

Pilot 5

HDRS digital governance and transaction management platform

This pilot would target the fragmented governance landscape, building and testing a unified digital layer where researchers apply for access, data controllers process requests, and all parties track projects through defined stages.

Unlike other pilots that test data and analytics capabilities, this pilot tests whether governance itself can become infrastructure. Such a pilot would directly operationalise Opportunity 5, creating the service layer through which all users interact with HDRS, regardless of which underlying data assets they require.

This aims to solve pervasive problems in meeting research delivery timelines. These include the need to process duplicative approvals and paperwork, difficulty in keeping track of projects where progress is a function of email chains and spreadsheets, and lack of transparency, auditability, and accountability. Practically, this means implementing portable researcher credentials recognised across participating data controllers, machine-readable templates for common agreement types, and transparent lifecycle tracking with published performance metrics.

Success would be measured by demonstrable reduction in time-to-data for projects using the platform, researcher adoption of portable credentials, and data controller adoption of standardised templates.

User stories served: all user stories

HDRS capabilities demonstrated: simpler access

Where to start: Limiting initial scope could de-risk development of what would be a novel platform that requires substantial process change. This could include covering only a subset of data controllers holding high-value assets, and focusing on a more narrow use-case (for example, NICE Health Technology Assessments, which require similar datasets and have time sensitive deadlines). Existing tools and processes for tackling some of these problems could be adapted, and those in active use by research organisations should be evaluated for this purpose. Success criteria could be used to gate further adoption across all HDRS activities.

Pilot 6

Near-real time device surveillance platform

This pilot would establish infrastructure for continuous safety surveillance of medical devices, with data flows operating at velocities appropriate for signal detection rather than retrospective research. The platform would incorporate device usage and safety outcomes extracted from clinical records, linked to structured data on prescribing, procedures, and clinical events.

Regulatory use cases have specific requirements: coverage must be sufficient for rare event detection, latency must support timely intervention, and data provenance must withstand regulatory scrutiny. Current UK infrastructure does not reliably meet these requirements. Device audit data is fragmented across registries, and missing where there are no manual collections. Safety signals in clinical text are not systematically extracted.

This targets Opportunities 1 and 2, but operationalises these for a use-case at the cutting edge of requirements, including testing whether source data engineering can be configured for surveillance, and whether an integration layer can support the data velocity required.

Success would be measured by regulatory use of the platform for actual safety assessments, demonstrating that HDRS can serve regulator requirements alongside research and commercial applications.

User stories served: regulator, pharma customer, clinical trial sponsor, medtech/biotech

HDRS capabilities demonstrated: opening up advanced diagnostics data

Where to start: This would be a technically ambitious pilot and could be de-risked by initially limiting scope to specialist centres that have existing advanced data infrastructure in place. Focusing on well-understood domains, where high quality UK-wide registry data collections already exist, could be used as a way to compare performance of existing methods with automated platform insights, to demonstrate that the latter can deliver comparable results.

Conclusion

This analysis has examined the UK health data ecosystem across all four nations, assessing the technical infrastructure, data assets, and non-technical constraints that will shape what the HDRS can achieve.

The review distinguished the technical requirements of scientific discovery and clinical research from those of healthcare operations. While operational systems focus on real-time care delivery, a world-class clinical research ecosystem requires a fundamentally different architecture: one built for high-dimensional data, longitudinal depth, multimodal integration, and the rigorous trust and governance standards demanded by patients, global trial sponsors, and clinical/academic researchers.

The landscape is heterogeneous. At every level, from nation to individual provider, data is generated, stored, processed, and made available through different systems, standards, and arrangements. This heterogeneity has allowed diverse architectures and networks to co-exist, developed organically in response to local relationships and requirements rather than national coordination. The result is a system where world-class components exist but cannot easily be combined to serve research needs requiring breadth, depth, and cross-nation coverage. Despite this, a number of organisations have demonstrated that high-quality, research-ready data can be delivered at scale, and with commercial viability. To scale this approach to a UK-wide research data service will be a challenge – but technological opportunities are achievable, and constraints, while substantial, are navigable with sustained attention.

What distinguishes the current moment is alignment. Government policy prioritises health data for research, economic growth, and AI development. Researcher demand for UK health data access is documented and growing. The technical patterns required for integration are maturing. Public and NHS support for research use of data, including with industry partners, remains strong where purpose is clear and safeguards are credible.

Without a clearly defined role that is focused on coordination rather than control, the HDRS will continue to face the same challenges identified in this report, regardless of investment or local improvement. The HDRS will be judged not by the elegance of its architecture but by whether researchers can access the data they need, when they need it, to answer questions that matter. The opportunities and pilots described in this report are only some possible pathways toward that outcome. The prioritisation of pathways, and their sequencing against available resources and organisational capacity, are decisions for the HDRS leadership to make.

Irrespective of what the final HDRS architecture will be, success will require attention to standardisation and integration, maintenance of trust and social licence, and bringing data custodians across all four nations into a shared endeavour – of a kind not previously achieved in the UK, but one with immense transformational potential.

Glossary

The following definitions are given to enable consistent and precise discussion throughout this review. They are not intended to create a comprehensive taxonomical coverage, nor to create canonical definitions outside of this review.

Ad Hoc Data Transfer

In contrast to these structured approaches, much of the current UK landscape currently operates through ad hoc arrangements: uncoordinated point-to-point data transfers – including by email – and manual processes that do not conform to any consistent architectural pattern. Such arrangements are often functional for individual projects, but they do not scale, cannot be audited consistently, and create cumulative governance burden and risk across the system.

Data Asset

A collection of person-level health-related data of any type, held in a single architecture and managed by a single organisational entity.

Data Asset Integration

The process of creating a single view over data assets representing different populations from separate locations. For example, if similar data from the populations of different hospital trusts were brought into a single TRE for analysis. Four approaches to integrating data assets are defined below, including Persistent Replication, Federated Analytics, Federated Learning, and the Data Fabric.

Data Discovery

The technology components and processes that enable end-users to understand what data of a specific description exists, and where it is available for access, across numerous data assets.

Data Engineering

The use of programming code (e.g., SQL, Python) to transform data from raw forms into the form needed for consumption by an end-user or a process. Data engineering may occur before and/or after data ingestion into a data asset.

Data Fabric

An architectural pattern for orchestrating separate data assets via metadata catalogues, and analysis of multiple assets from different access points. A fabric addresses the question: "How do I make data across an ecosystem accessible and analysable regardless of where it resides?" Methods include generating virtual views over datasets and coordinated temporary replication for single analyses, but can also include orchestration of 'eyes off' federated approaches.

Data Linkage

The process of combining data from different sources for a single individual. For example, linking a person's primary care record to their secondary care Hospital Episode Statistics. Linkage encompasses both the matching process (typically using NHS number or other persistent identifiers) and the technological methods for bringing data of different provenances into the same environment.

Data Management

The database or data lake architecture that stores and organises data within an asset. This encompasses storage infrastructure and the technical operations required to maintain data availability and integrity.

Data Model

A specification defining the syntactic (table structure) and semantic (concept meaning) representation of data. Common data models (such as the Observational Medication Outcomes Partnership, or OMOP) represent a standardisation layer enabling consistent representation across assets.

Data Provenance

The documented source of data and all transformations it has undergone. Provenance is essential for scientific work, particularly in life sciences, as it captures contextual meaning that may not be apparent from the data itself (e.g., an outpatient diagnostic code carries different clinical implications from an inpatient code for the same condition).

Federated Analytics

Query-time aggregation and integration of data across separate sources without moving the underlying data. Federated approaches are usually synonymous with 'eyes off', with analysis code being sent to data locations, executed locally, and only aggregated results being returned. This method relies on data at source having a common data model (sharing table structure and concepts), or on queries undergoing specific 'translation' to match the source data model. Robust analytics also relies on data at different sources having consistent provenance.

Federated Learning

A machine learning approach where models are trained across decentralised data sources without physically transferring the underlying data. Rather than aggregating

query results, federated learning shares model parameters (such as weights or gradients) which are aggregated centrally to update a global model. This enables training on larger and more diverse populations than any single data holder could provide, while keeping what is often 'big' data in place. Federated learning requires computational resources to be located where the data is held.

Persistent Replication

The copying of data assets to a new location where they are persisted as a new asset. This definition purposefully excludes cases where data is temporarily replicated to enable a single analysis followed by deletion of the replica once work is complete.

Trusted Research Environment (TRE)

Any secure platform that manages approved user access to data while conforming to the Five Safes framework. This review considers any UK platform for secure health data access to be a TRE, not restricted to those with ISO 27001 certification, and including Data Safe Havens and Secure Data Environments. Note that in many existing cases, research data analysis takes place outside TREs via a physical dataset transfer to end-user systems.

Acknowledgements and attributions

This independent report was commissioned by the Wellcome Trust. The authors would like to acknowledge the valuable contributions of a range of stakeholders who provided expert input to the landscape analysis and review of the HDRS design. Contributions were provided through interviews, workshops, and written feedback, and informed the authors' understanding of strengths, challenges, and limitations within current health data ecosystems. Stakeholder insights also helped to shape the consideration of the pilot approaches presented in this report.

Individual contributors

We are grateful to the following individuals who contributed their expertise to this report:

- Albert King, NHS National Services Scotland, Chief Data Officer
- Claire Bloomfield, Isomorphic Labs
- Claire MacDonald, Clinical Data Science Unit, Manchester University NHS Foundation Trust, Head of Clinical Data Science
- Daniel Prieto-Alhambra, University of Oxford, Lead, Health Data Sciences
- Dr Guy Tsafnat, Evidentli, Founder, Chief Scientific Officer
- David James, Prostate Cancer Research, Director
- Dr Kristin-Anne Rutter, Cambridge University Health Partners
- Dr Maria Koufali, NIHR, Life Sciences Industry Director
- Dr Nicole Mather, IBM UKI & EMEA, Life Sciences, Partner
- Felix Ritchie, University of the West of England Bristol, Director
- James Fleming, Fraktal Consulting Ltd, Managing Principal
- James McCafferty, Wellcome Sanger Institute, Chief Information Officer
- James Peach, Bio Industry Association, Data Access Committee
- Jonathan Smart, SAIL Databank and the Secure eResearch Platform (SeRP), Chief Operating Officer
- Karen Ambrose, Francis Crick Institute, Chief Data Officer
- Layla Robinson, Research Data Scotland, Chief Operating Officer
- Lill-Brith Wium von Arx, Eli Lilly and Company, Senior Director, Health Innovation & Evidence Lead, Northern Europe
- Mark Avery, Cambridge University Health Partners (CUHP) and Health Innovation East, Eastern England Secure Data Environment Director
- Prof Martin Gibson, NWEH Ltd, Chief Medical Officer.
- Prof Matthew Brookes, NIHR RDN and Royal Wolverhampton NHS Trust, Regional Network Director West Midlands and Consultant Gastroenterologist
- Oliver Lake, Professional Record Standards Body, Chief Executive Officer
- Pearse Keane, Moorfields Eye Hospital, Director, INSIGHT Health Data Research Hub
- Prof Darren Treanor, National Pathology Imaging Cooperative, Director
- Prof Simon de Lusignan, University of Oxford, Professor of Primary Care and Clinical Informatics, Director Royal College of General Practitioner Research and Surveillance Centres
- Prof Iain E Buchan, University of Liverpool / Civic Health Innovation Labs, Associate Pro Vice Chancellor for Innovation and W.H. Duncan Professor of Public Health Systems
- Richard Walls, University of Dundee, Health Informatics Centre Operational Director
- Sally Stewart, North West Secure Data Environment, Director
- Tomas Sanchez Lopez, NHS England, Director Technology and Data Integration
- Vishnu V Chandrabalan, Lancashire and South Cumbria Secure Data Environment, Chief Clinical Information Officer

Organisational contributors

We are grateful to the following organisations who contributed expertise through representatives who requested organisational attribution only, with particular thanks to ABPI and TechUK for coordinating industry responses:

- Alzheimer's Society
- Association of the British Pharmaceutical Industry (ABPI)

- BioIndustry Association (BIA)
- Cambridge University Health Partners
- DataLoch
- Eastern England Secure Data Environment
- Francis Crick Institute
- Genomics Ltd
- Health and Social Care Northern Ireland (HSCNI)
- Health Data Research UK (HDR UK)
- Imperial College London
- INSIGHT Health Data Research Hub
- IQVIA
- Lancashire and South Cumbria Secure Data Environment
- LifeArc
- Manchester University NHS Foundation Trust
- Medical Research Council (MRC)
- Medicines and Healthcare products Regulatory Agency (MHRA)
- National Pathology Imaging Cooperative (NPIC)
- NHS England
- NHS Greater Manchester
- NHS National Services Scotland
- NIHR Policy Research Unit in Reproductive Health
- North West Region Secure Data Environment
- NWEH Ltd
- Our Future Health
- PHG Foundation
- Professional Record Standards Body (PRSB)
- Prostate Cancer Research
- Public Health Scotland
- Relation Therapeutics
- Research Data Scotland
- Resolution Therapeutics
- SAIL Databank
- Secure eResearch Platform (SeRP)
- Sanofi UK & Ireland
- Scottish Government
- Starlight Consulting
- techUK
- Thames Valley and Surrey Secure Data Environment
- Medical Research Council
- UK Longitudinal Linkage Collaboration
- UK Space Agency
- University of Dundee
- University of Edinburgh
- University of Liverpool
- University of Manchester
- University of Oxford
- University of the West of England Bristol
- Wellcome Sanger Institute
- Yorkshire and Humber Secure Data Environment

The views expressed in this independent report are those of the authors and do not necessarily reflect the views of the individuals or organisations acknowledged. We thank all contributions to this report, not all of which can be acknowledged. All errors and omissions remain the responsibility of the authors.

About the authors

This paper was authored by members of the Emrys Health Consortium, commissioned by the Wellcome Trust to deliver an independent landscape review of the UK's health data infrastructure and to develop recommendations and pilot approaches to inform the design of the HDRS.

The consortium comprises Emrys Health Ltd, Nesta (the UK's innovation charity), Newmarket Strategy, Yale University, supported by individual expert input.

Contributors from Emrys Health Ltd include Pollyanna Jones and Will Browne (Co-founders), Ben Levinson (Director), and Emily Jones (Associate). Contributors from Nesta include Mallory Durran (Applied Research and Methods Director), Will Woodward (Mission Discovery Lead), and Natalie Lai (Senior Analyst for Mission Discovery). Contributors from Newmarket Strategy included Ele Harwich (Director). Contributors from Yale University included Dr Jess Morley (Associate Research Scientist). Dr Joe Zhang (CTO of the AI Centre for Value-Based Healthcare) contributed as an individual data and technology expert.

Collectively, the authors bring experience spanning government, academia, industry, and the third sector, with a shared focus on enabling trustworthy, effective, and impactful use of health data for research and public benefit.

Authors

- Will Browne
- Mallory Durran
- Ele Harwich
- Pollyanna Jones
- Natalie Lai
- Ben Levinson
- Dr Jess Morley
- Will Woodward
- Dr Joe Zhang

References

1. Department for Science, Innovation and Technology. AI Opportunities Action Plan. https://www.gov.uk/government/publications/ai-opportunities-action-plan (2025).
2. Department of Health and Social Care & Prime Minister's Office, 10 Downing Street. Fit for the Future: 10 Year Health Plan for England. https://www.gov.uk/government/publications/10-year-health-plan-for-england-fit-for-the-future/fit-for-the-future-10-year-health-plan-for-england-accessible-version (2025).
3. Department for Business and Trade, Department of Health & Social Care, Department for Science, Innovation and Technology & Office for Life Sciences. Life Sciences Sector Plan. https://www.gov.uk/government/publications/life-sciences-sector-plan (2025).
4. Sudlow, C. Uniting the UK's Health Data: A Huge Opportunity for Society. https://www.hdruk.ac.uk/helping-with-health-data/the-sudlow-review (2024).
5. Goldacre, B., Morley, J. & Department of Health and Social Care. Better, Broader, Safer: Using Health Data for Research and Analysis. https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis (2022).
6. Wachter, R. M. Making IT Work: Harnessing the Power of Health Information Technology to Improve Care in England. Making IT Work 71 (2016).
7. O'Shaughnessy, J., Department for Health and Social Care, Department for Science, Innovation and Technology & Office for Life Sciences. Commercial Clinical Trials in the UK: The Lord O'Shaughnessy Review – Final Report. https://www.gov.uk/government/publications/commercial-clinical-trials-in-the-uk-the-lord-oshaughnessy-review/commercial-clinical-trials-in-the-uk-the-lord-oshaughnessy-review-final-report (2023).
8. NHS England » NHS Federated Data Platform infrastructure. NHS England https://www.england.nhs.uk/digitaltechnology/nhs-federated-data-platform/nhs-fdp-explained/how-does-the-nhs-federated-data-platform-work/ (2026).
9. HDR UK. The NHS Research Secure Data Environment (SDE) Network. https://healthdatagateway.org/en/data-custodian-network/3 (2026).
10. Brophy, R. et al. Towards a standardised cross-sectoral data access agreement template for research: a core set of principles for data access within trusted research environments. International Journal of Population Data Science 8, (2023).
11. Edwards, L. et al. UK research data resources based on primary care electronic health records: review and summary for potential users. BJGP Open 7, BJGPO.2023.0 (2023).
12. UK Biobank. About our data. UK Biobank https://www.ukbiobank.ac.uk/about-our-data/ (2025).
13. Genomics England. Bioinformatics and data science. Genomics England https://www.genomicsengland.co.uk/bioinformatics (2026).
14. Our Future Health. How Our Future Health works. Our Future Health https://ourfuturehealth.org.uk/our-research-mission/how-our-future-health-works/ (2026).
15. McDonald, P. L., Phillips, J., Harwood, K., Maring, J. & Van Der Wees, P. J. Identifying requisite learning health system competencies: a scoping review. BMJ Open e061 (2022).
16. Hardie, T., Horton, T., Thornton-Lee, N., Home, J. & Pereira, P. Developing Learning Health Systems in the UK: Priorities for Action. https://www.health.org.uk/publications/reports/developing-learning-health-systems-in-the-uk-priorities-for-action
17. Velummailum, R. R. et al. Data Challenges for Externally Controlled Trials: Viewpoint. J Med Internet Res e43 (2023).
18. Jamie, G. et al. Phenotype execution and modeling architecture to support disease surveillance and real-world evidence studies: English sentinel network evaluation. JAMIA Open 7, ooae (2024).
19. Dang, A. Real-World Evidence: A Primer. Pharm Med 25–36 (2023).
20. Department for Science, Innovation and Technology & Office for Artificial Intelligence. A Pro-Innovation Approach to AI Regulation. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper (2023).
21. Arora, A. et al. The value of standards for health datasets in artificial intelligence-based applications. Nat Med 2929–2 (2023).
22. World Health Organization. Generating Evidence for Artificial Intelligence-Based Medical Devices: A Framework for Training, Validation and Evaluation. (World Health Organization, Geneva, 2021).
23. Morley, J. On designing an algorithmically enhanced NHS: towards a conceptual model for the successful implementation of algorithmic clinical decision support software in the National Health Service. (University of Oxford, 2023).
24. Shabani, M. Will the European Health Data Space change data sharing rules? Science 31357–1 (2022).
25. Terzis, P. Compromises and Asymmetries in the European Health Data Space. Eur. J. Health Law 345 (2022).
26. El Sabawy, D., Feldman, J. & Pinto, A. D. The Connected Care for Canadians Act : an important step toward interoperability of health data. CMAJ 1E1385–E1 (2024).
27. ABPI. UK Industry Clinical Trials: Translating Actions into Impact. https://www.abpi.org.uk/publications/uk-industry-clinical-trials-translating-actions-into-impact/ (2025).
28. Department of Health and Social Care, Prime Minister's Office, 10 Downing Street, Department for Science, Innovation and Technology, Office for Life Sciences & Office for Investment. Prime Minister turbocharges medical research. GOV.UK https://www.gov.uk/government/news/prime-minister-turbocharges-medical-research (2025).
29. DARE UK. DARE UK: About us. DARE UK https://dareuk.org.uk/about-us/ (2026).
30. Department of Health and Social Care. The future of healthcare: our vision for digital, data and technology in health and care. GOV.UK https://www.gov.uk/government/publications/the-future-of-healthcare-our-vision-for-digital-data-and-technology-in-health-and-care/the-future-of-healthcare-our-vision-for-digital-data-and-technology-in-health-and-care (2018).
31. Department of Health and Social Care. Data saves lives: reshaping health and social care with data. GOV.UK https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data (2022).
32. Pawson, R. Evidence-based Policy: The Promise of `Realist Synthesis'. Evaluation 8, 340 (2002).
33. Pawson, R., Greenhalgh, T., Harvey, G. & Walshe, K. Realist review – a new method of systematic review designed for complex policy interventions. J Health Serv Res Policy 21–34 (2005).
34. Heeks, R., Mundy, D. & Salazar, A. Why Health Care Information Systems Succeed or Fail. (Institute for Development Policy and Management, Manchester, 1999).
35. Griffiths, E. et al. Findability of UK health datasets available for research: A mixed methods study. BMJ Health and Care Informatics (2022).
36. DICOM Standard. About DICOM- Overview. DICOM https://www.dicomstandard.org/about (2026).

37. NIH. File Format Guide: National Center for Biotechnology Information. National Library of Medicine https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/.

38. HL7. Introducing HL7 FHIR. https://hl7.org/fhir/summary.html (2023).

39. The OpenEHR Foundation. OpenEHR: Architecture Overview. https://specifications.openehr.org/releases/BASE/latest/architecture_overview.html (2021).

40. OHDSI. Standardized Data: The OMOP Common Data Model. https://www.ohdsi.org/data-standardization/ (2026).

41. SNOMED. What is SNOMED CT. SNOMED International https://www.snomed.org/what-is-snomed-ct (2026).

42. NHS England. National Clinical Coding Standards ICD-10 5th Edition for Morbidity Coding (2025). https://classbrowser.nhs.uk/#/book/ICD-10-5TH-Edition (2025).

43. NHS England. National Clinical Coding Standards OPCS-4 (2025). https://classbrowser.nhs.uk/#/book/OPCS-4.10 (2025).

44. NHS Business Services Authority. About dm+d. https://www.nhsbsa.nhs.uk/pharmacies-gp-practices-and-appliance-contractors/nhs-dictionary-medicines-and-devices-dmd/about-dmd (2026).

45. Vorisek, C. N. et al. Towards an Interoperability Landscape for a National Research Data Infrastructure for Personal Health Data. Sci Data (2024).

46. Torab-Miandoab, A., Samad-Soltani, T., Jodati, A. & Rezaei-Hachesu, P. Interoperability of heterogeneous health information systems: a systematic literature review. BMC Med Inform Decis Mak 18 (2023).

47. Zhang, J., Ashrafian, H., Delaney, B. & Darzi, A. Impact of primary to secondary care data sharing on care quality in NHS England hospitals. npj Digit. Med. 6, (2023).

48. Lee, S. et al. Unlocking the Potential of Electronic Health Records for Health Research. IJPDS 5, (2020).

49. Ntinopoulos, V. et al. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. BMJ Health Care Inform e101(2025).

50. Harron, K. Data linkage in medical research. bmjmed 1, e000(2022).

51. UK Biobank. Research Analysis Platform. UK Biobank https://www.ukbiobank.ac.uk/use-our-data/research-analysis-platform/ (2025).

52. Liddell, K., Simon, D. A. & Lucassen, A. Patient data ownership: who owns your health? Journal of Law and the Biosciences 8, Isab (2021).

53. Igbo, W. Understanding the TRE landscape: Early insights from DARE UK's 2infrastructure landscape review. DARE UK https://dareuk.org.uk/news-and-events/understanding-the-tre-landscape-early-insights-from-dare-uks-2025-infrastructure-landscape-review/ (2025).

54. Kavianpour, S., Sutherland, J., Mansouri-Benssassi, E., Coull, N. & Jefferson, E. Next-Generation Capabilities in Trusted Research Environments: Interview Study. Journal of Medical Internet Research (2022).

55. Casaletto, J., Bernier, A., McDougall, R. & Cline, M. S. Federated Analysis for Privacy-Preserving Data Sharing: A Technical and Legal Primer. Annu. Rev. Genom. Hum. Genet. 347 (2023).

56. Nițulescu, A. & Stoicu-Tivadar, L. Data Standardization in the Medical Field Through FHIR and FAIR Implementation: A Systematic Review in Studies in Health Technology and Informatics (eds Mantas, J. et al.) (IOS Press, 2024). doi:10.3233/SHTI2406

57. NHS England. Master Person Service (MPS). NHS England Digital https://digital.nhs.uk/services/personal-demographics-service/master-person-service (2026).

58. Office for Statistics Regulation. Data Sharing and Linkage for the Public Good. https://osr.statisticsauthority.gov.uk/wp-content/uploads/2023/07/202307_office_statistics_regulation_data_sharing_linkage_report.pdf (2023).

59. Public Health Scotland. About the CHI linkage and indexing team (CHILI). https://publichealthscotland.scot/resources-and-tools/health-intelligence-and-data-management/chi-linkage-and-indexing-chili/about-the-chi-linkage-and-indexing-team-chili/ (2026).

60. John Snow Labs. State of the Art Clinical Data Curation. John Snow Labs https://www.johnsnowlabs.com/state-of-the-art-clinical-data-curation/ (2026).

61. CogStack. Unlock the Power of Healthcare Data with CogStack. Cogstack https://cogstack.org/ (2023).

62. Akrivia Health. Akrivia Health. Akrivia Health https://akriviahealth.com/.

63. AI Centre for Value Based Healthcare. Large Language Models ('OncoLlama') for Pan-Cancer Data Enrichment. AI Centre https://www.aicentre.co.uk/news-and-events/news/large-language-models-oncollama-for-pan-cancer-data-enrichment.

64. Reich, C. et al. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. Journal of the American Medical Informatics Association 583 (2024).

65. Reinecke, I., Zoch, M., Reich, C., Sedlmayr, M. & Bathelt, F. The Usage of OHDSI OMOP – A Scoping Review. in Studies in Health Technology and Informatics (eds Röhrig, R. et al.) (IOS Press, 2021). doi:10.3233/SHTI2105

66. González-García, J. et al. PHIRI: lessons for an extensive reuse of sensitive data in federated health research. European Journal of Public Health i43–i49 (2024).

67. HDR UK. Cohort Discovery. Health Data Gateway https://healthdatagateway.org/en/about/cohort-discovery (2026).

68. Sebire, N. J., Cake, C. & Morris, A. D. HDR UK supporting mobilising computable biomedical knowledge in the UK. BMJ Health Care Inform e100 (2020).

69. Tibble, H. et al. Using routine primary care data in research: (in)efficient case studies and perspectives from the Asthma UK Centre for Applied Research. BMJ Health Care Inform e101(2025).

70. Nab, L. et al. OpenSAFELY: a platform for analysing electronic health records designed for reproducible research. Preprint at https://doi.org/10.31219/osf.io/hj2sg (2024).

71. Gaye, A. et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. International Journal of Epidemiology 1929–1(2014).

72. Lifebit. Data Intelligence Platform – https://lifebit.ai/ (2025).

73. Bitfount. Federated AI and Data Science Platform. https://www.bitfount.com (2026).

74. FITFILE https://fitfile.com/ (2026).

75. TriNetX. Real-world data for the life sciences and healthcare. TriNetX https://trinetx.com/ (2026).

76. Parimi, S. Data Mesh vs. Data Fabric: The Future of Data Management. IJSAT 2(2025).

77. Nathan, M. & Srinivasan, S. A Fresh Look: The Role of a Healthcare Data Fabric in AI-Driven Predictions. IJSAT 8(2025).

78. Zhang, P. & Kamel Boulos, M. N. Privacy-by-Design Environments for Large-Scale Health Research and Federated Learning from Data. IJERPH 11(2022).

79. Dr Christian Cole et al. SATRE: Standardised Architecture for Trusted Research Environments. https://zenodo.org/doi/10.5281/zenodo.10055(20doi:10.5281/ZENODO.100553

80. DNAnexus®. The Precision Health Data Cloud. https://www.dnanexus.com (2026).

81. Walsh, L. The rise of Dawn. https://www.cam.ac.uk/stories/ai-supercomputer-dawn-research-energy-medicine-climate (2024).

82. EPCC. Delivering UK supercomputing and data science excellence to the world. https://www.epcc.ed.ac.uk/ (2026).

83. University of Bristol. Bristol Centre for Supercomputing (BriCS). https://www.bristol.ac.uk/research/centres/bristol-supercomputing/ (2026).

84. DARE UK. FRIDGE: Federated Research Infrastructure by Data Governance Extension. DARE UK https://dareuk.org.uk/how-we-work/ongoing-activities/dare-uk-early-adopters/fridge/ (2026).

85. Gierend, K. et al. Provenance Information for Biomedical Data and Workflows: Scoping Review. J Med Internet Res e51(2024).

86. Government Digital Service, Department for Science, Innovation, and Technology. Guidelines and best practices for making government datasets ready for AI. GOV.UK https://www.gov.uk/government/publications/making-government-datasets-ready-for-ai/guidelines-and-best-practices-for-making-government-datasets-ready-for-ai (2026).

87. Green, E., Kendal, C., Ritchie, F. & Alves, K. Setting up Output Checking Processes: A Guide for Data Services. (2024).

88. Alves, K. & Ritchie, F. Runners, repeaters, strangers and aliens: Operationalising efficient output disclosure control. SJI 1281–1 (2020).

89. Mansouri-Benssassi, E. et al. Disclosure control of machine learning models from trusted research environments (TRE): New challenges and opportunities. Heliyon 9, e15 (2023).

90. HDR UK. Safe People Registry. Safe People Registry https://safepeopleregistry.org/en/about (2025).

91. MHRA. Archiving and retention of clinical trial records. GOV.UK https://www.gov.uk/government/publications/archiving-and-retention-of-clinical-trial-records/archiving-and-retention-of-clinical-trial-records (2026).

92. Gavin, K. M., Sundermann, M. L. & Wieland, A. Leveraging real-world data for safety signal detection and risk management in pre- and post-market settings. Front. Drug Saf. Regul. 5, 1626 (2025).

93. Zhang, J. et al. Mapping and evaluating national data flows: transparency, privacy, and guiding infrastructural transformation. The Lancet Digital Health 5, (2023).

94. Nab, L. et al. OpenSAFELY : A platform for analysing electronic health records designed for reproducible research. Pharmacoepidemiology and Drug e5 (2024).

95. NHS England. DigiTrials. NHS England Digital https://digital.nhs.uk/services/nhs-digitrials (2025).

96. NWEH. NWEH: Our mission. NWEH https://www.nweh.co.uk/about/our-mission/ (2026).

97. NWEH. FARSITE. NWEH https://www.nweh.co.uk/platforms/farsite/ (2026).

98. NWEH. ConneXon. NWEH https://www.nweh.co.uk/platforms/connexon/ (2026).

99. Wirth, F. N., Meurers, T., Johns, M. & Prasser, F. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. BMC Med Inform Decis Mak (2021).

100. Dataloch. Dataloch. https://dataloch.org/ (2026).

101. University of Dundee. Health Informatics Centre (HIC). https://www.dundee.ac.uk/hic (2026).

102. The University of Aberdeen. Safe Haven (DaSH). https://www.abdn.ac.uk/research/digital/platforms/safe-haven-dash/ (2026).

103. Baxter, R. et al. The Scottish Medical Imaging Archive: 57.3 Million Radiology Studies Linked to Their Medical Records. Radiology: Artificial Intelligence 6, e220 (2024).

104. Camilleri, M. P. J. et al. A large dataset of brain imaging linked to health systems data: the curation and access to a whole system national cohort from NHS Scotland. Preprint at https://doi.org/10.1101/2025.10.21.25338 (2025).

105. HSC Business Services Organisation. Honest Broker Service. Business Services Organisation (BSO) Website https://bso.hscni.net/directorates/digital/honest-broker-service/ (2026).

106. Digital Health and Care Northern Ireland. Analytics & Insight. DHCNI https://dhcni.hscni.net/digital-strategy/data/analytics_insight/ (2026).

107. Boyd, A. et al. UK Longitudinal Linkage Collaboration (UK LLC): The National Trusted Research Environment for Longitudinal Research. IJPDS (2025).

108. NIHR BioResource. NIHR BioResource. https://bioresource.nihr.ac.uk/ (2026).

109. UK Biobank. UK BioBank: Primary Care Linked Data. (2024).

110. NIHR. Biomedical Research Centres. https://www.nihr.ac.uk/about-us/what-we-do/infrastructure/biomedical-research-centres (2026).

111. IQVIA. Real World & Healthcare Data. https://www.iqvia.com/solutions/real-world-evidence/real-world-data-and-insights (2026).

112. Flatiron Health. Reimagining the infrastructure of cancer care. https://flatiron.com (2026).

113. Arcturis. Advancing Insights Using Real-World Data. Arcturis https://www.arcturisdata.com/ (2026).

114. DARE UK (Data and Analytics Research Environments UK). DARE UK Federated Architecture Blueprint. https://doi.org/10.5281/ZENODO.14192 (2024).

115. Avraam, D. et al. DataSHIELD: mitigating disclosure risk in a multi-site federated analysis platform. Bioinformatics Advances 5, vbaf (2024).

116. Martin, F. et al. vantage6. Zenodo https://doi.org/10.5281/ZENODO.7221 (2026).

117. Flower AI. Flower Framework Documentation. Flower https://flower.ai/docs/framework/index.html (2026).

118. AI Centre for Value Based Healthcare. Our Platforms. AI Centre https://www.aicentre.co.uk/our-platforms (2026).

119. NHS England. Hospital Episode Statistics (HES). NHS England Digital https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics (2025).

120. Public Health Scotland. What are the SMR datasets? – Scottish Morbidity Records (SMR) – Data management in secondary care: hospital activity – Health intelligence and data management – Resources and tools – Public Health Scotland. https://publichealthscotland.scot/resources-and-tools/health-intelligence-and-data-management/data-management-in-secondary-care-hospital-activity/scottish-morbidity-records-smr/what-are-the-smr-datasets/ (2024).

121. Health Data Research Gateway. Patient Episode Dataset for Wales (PEDW). https://healthdatagateway.org/en/dataset/ (2025).

122. SAIL Databank. SAIL Databank is a rich and trusted population databank. SAIL Databank https://saildatabank.com/ (2025).

123. Discover-NOW. Health Data Research Hub for RWE I Discover-NOW. Discover Now https://discover-now.co.uk/ (2026).

124. NHS England. Linked HES-ONS mortality data. NHS England Digital https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/linked-hes-ons-mortality-data (2023).

125. NHS England. NHS OpenSAFELY Data Analytics Service Pilot Directions 20NHS England Digital https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/directions-and-data-provision-notices/secretary-of-state-directions/nhs-opensafely-data-analytics-service-pilot-directions-2 (2026).

126. Lehmann, B. et al. Methodological opportunities in genomic data analysis to advance health equity. Nat Rev Genet 635 (2025).

127. Al-Sahab, B., Leviton, A., Loddenkemper, T., Paneth, N. & Zhang, B. Biases in Electronic Health Records Data for Generating Real-World Evidence: An Overview. J Healthc Inform Res 8, 121 (2024).

128. The RECOVERY Collaborative Group. Dexamethasone in Hospitalized Patients with Covid-N Engl J Med 3693 2021).

129. NHS-Galleri Trial. NHS-Galleri Trial: Detecting cancer early. NHS-Galleri Trial https://www.nhs-galleri.org/ (2021).

130. Hou, J. et al. Generate Analysis-Ready Data for Real-world Evidence: Tutorial for Harnessing Electronic Health Records With Advanced Informatic Technologies. J Med Internet Res e45 (2023).

131. Reilly, G. & Varma, S. Health Data Research Innovation Gateway. ITNOW 60–63 (2021).

132. Public Health Scotland. Electronic Data Research and Innovation Service (eDRIS). https://publichealthscotland.scot/resources-and-tools/health-intelligence-and-data-management/electronic-data-research-and-innovation-service-edris/overview/ (2025).

133. Office for National Statistics. Secure Research Service. https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice (2026).

134. Scottish Parliament. Care Reform (Scotland) Act 20https://www.legislation.gov.uk/asp/2025/9 (2026).

135. Harron, K., Doidge, J. C. & Goldstein, H. Assessing data linkage quality in cohort studies. Annals of Human Biology 218 (2020).

136. Department of Health and Social Care. Data Saves Lives: Reshaping Health and Social Care with Data. https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data (2022).

137. National Audit Office. Challenges in using data across government – NAO insight. National Audit Office (NAO) https://www.nao.org.uk/insights/challenges-in-using-data-across-government/ (2019).

138. Department of Health & Social Care. Putting Data, Digital and Tech at the Heart of Transforming the NHS. https://www.gov.uk/government/publications/putting-data-digital-and-tech-at-the-heart-of-transforming-the-nhs/putting-data-digital-and-tech-at-the-heart-of-transforming-the-nhs (2021).

139. Ameen, S., Wong, M. C., Yee, K. C., Nøhr, C. & Turner, P. The Inverse Data Law: Market Imperatives, Data, and Quality in AI Supported Care. in Studies in Health Technology and Informatics (eds Hägglund, M. et al.) (IOS Press, 2023). doi:10.3233/SHTI2301

140. Susser, D. et al. Synthetic Health Data: Real Ethical Promise and Peril. Hastings Center Report 8–13 (2024).

141. Taylor, J. A. et al. The road to hell is paved with good intentions: the experience of applying for national data for linkage and suggestions for improvement. BMJ Open e047 (2021).

142. Villacorta Linaza, R. et al. Overcoming barriers to NHS adoption of innovative IPC products: A qualitative study of SMEs in the Liverpool city region. PLoS One e0331 (2025).

143. Kerasidou, A. & Kerasidou, C. Data-driven research and healthcare: public trust, data governance and the NHS. BMC Med Ethics 51 (2023).

144. Tazare, J. et al. NHS National Data Opt-Outs: trends and potential consequences for health data research. BJGP Open BJGPO.2024.0 (20doi:10.3399/BJGPO.2024.00)

145. Byrne, N. National Data Guardian 2023-2Report. https://www.gov.uk/government/publications/national-data-guardian-2023-2024-report/national-data-guardian-2023-2024-report (2024).

146. Meszaros, J. & Ho, C. Building trust and transparency? Challenges of the opt-out system and the secondary use of health data in England. Medical Law International 159 (2019).

147. Horn, R. & Kerasidou, A. Sharing whilst caring: solidarity and public trust in a data-driven healthcare system. BMC Med Ethics (2020).

148. Jones, K. H. et al. The other side of the coin: Harm due to the non-use of health-related data. International Journal of Medical Informatics 43–51 (2017).

149. Sterckx, S., Rakic, V., Cockbain, J. & Borry, P. "You hoped we would sleep walk into accepting the collection of our data": controversies surrounding the UK care.data scheme and their wider relevance for biomedical research. Med Health Care and Philos 177 (2016).

150. Carter, P., Laurie, G. T. & Dixon-Woods, M. The social licence for research: why care.data ran into trouble. J Med Ethics 404 (2015).

151. Cumyn, A. et al. Patients' and Members of the Public's Wishes Regarding Transparency in the Context of Secondary Use of Health Data: Scoping Review. Journal of Medical Internet Research (2023).

152. Aitken, M. et al. Consensus Statement on Public Involvement and Engagement with Data-Intensive Health Research. International Journal of Population Data Science 4 (2019).

153. Bazzano, A. N., Mantsios, A., Mattei, N., Kosorok, M. R. & Culotta, A. AI Can Be a Powerful Social Innovation for Public Health if Community Engagement Is at the Core. J Med Internet Res e68 (2025).

154. Shaw, J. A., Sethi, N. & Cassel, C. K. Social license for the use of big data in the COVID-19 era. npj Digit. Med. 3 (2020).

155. Goldacre, B. et al. Bringing NHS data analysis into the 21st century. J R Soc Med 1383 (2020).

156. UKRI. Research financial sustainability: issues paper. https://www.ukri.org/publications/research-financial-sustainability-data/research-financial-sustainability-issues-paper/ (2023).

157. Government Commercial Function. The Procurement Act 20A short guide for suppliers. GOV.UK https://www.gov.uk/government/publications/procurement-act-2023-short-guides/the-procurement-act-2023-a-short-guide-for-suppliers-html (2025).

158. Federated Research Technical Documentation, Five SAFES TES https://docs.federated-analytics.ac.uk/five_safes_tes (2026)

159. Aridhia https://www.aridhia.com/ (2025).

160. BC Platforms Accelerate innovation and outcomes https://www.bcplatforms.com/ (2025)

161. Pineda-Moncusí, M., Rahman, M., Axson, E.L. et al. Academic impact and research data utilisation of the clinical practice research datalink: scientometric analyses. Eur J Epidemiol (2026).

162. Mueller, T et al, Data Resource Profile: The Scottish Combined Medicines Dataset (SCoMeD) International Journal of Population Data Science (208:6:14)

163. EDGE: Research Management in Real Time. https://edgeclinical.com/ (20

164. Schoeler. "Participation bias in the UK Biobank distorts genetic associations and downstream analyses." Nature human behaviour, vol. 7.7, 20pp. 1216-1227.

About Wellcome

Wellcome supports science to solve the urgent health challenges facing everyone. We support discovery research into life, health and wellbeing, and we're taking on three worldwide health challenges: mental health, global heating and infectious diseases.

Designed by Nesta for Wellcome.

Wellcome Trust, 215 Euston Road, London NW1 2BE, United Kingdom
T +44 (0)20 7611 8888, E contact@wellcome.org, wellcome.org

The Wellcome Trust is a charity registered in England and Wales, no. 210183.
Its sole trustee is The Wellcome Trust Limited, a company registered in England and Wales, no. 2711000
(whose registered office is at 215 Euston Road, London NW1 2BE, UK).