Mapping and evaluating national data flows: transparency, privacy, and guiding infrastructural transformation

Joe Zhang, Jess Morley, Jack Gallifant, Chris Oddy, James T Teo, Hutan Ashrafian, Brendan Delaney, Ara Darzi

The importance of big health data is recognised worldwide. Most UK National Health Service (NHS) care interactions are recorded in electronic health records, resulting in an unmatched potential for population-level datasets. However, policy reviews have highlighted challenges from a complex data-sharing landscape relating to transparency, privacy, and analysis capabilities. In response, we used public information sources to map all electronic patient data flows across England, from providers to more than 460 subsequent academic, commercial, and public data consumers. Although NHS data support a global research ecosystem, we found that multistage data flow chains limit transparency and risk public trust, most data interactions do not fulfil recommended best practices for safe data access, and existing infrastructure produces aggregation of duplicate data assets, thus limiting diversity of data and added value to end users. We provide recommendations to support data infrastructure transformation and have produced a website (https://DataInsights.uk) to promote transparency and showcase NHS data assets.

Lancet Digit Health 2023; 5: e737–48

Institute of Global Health Innovation, Imperial College London, London, UK (J Zhang BMBCh, Prof H Ashrafian PhD, Prof B Delaney PhD, Prof A Darzi PhD)**; Department of Critical Care** (J Zhang) **and London Medical Imaging and AI Centre** (Prof J T Teo PhD)**, Guy's and St Thomas' NHS Foundation Trust, London, UK; Oxford Internet Institute, University of Oxford, Oxford, UK** (J Morley MS)**; Department of Intensive Care, Imperial College Healthcare NHS Trust, London, UK** (J Gallifant MSc)**; Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA** (J Gallifant)**; Department of Anaesthesia, Critical Care and Pain, St George's Healthcare NHS Trust, London, UK** (C Oddy MBBS)**; Department of Neurology, King's College Hospital NHS Foundation Trust, London, UK** (Prof J T Teo)**; Leeds University Business School, Leeds, UK** (Prof H Ashrafian)

Correspondence to:
Dr Joe Zhang, Institute of Global Health Innovation, Imperial College London, London SW7 2AZ, UK
joe.zhang@imperial.ac.uk

Introduction

Digital transformation in the UK National Health Service (NHS) has resulted in most present and historical patient interactions being stored within electronic health record systems.[1] A focus on interoperability has enabled widespread data sharing between discrete systems, making medical records available for direct care, but also enabling aggregation of large datasets for secondary uses. As a result, data within NHS systems are a valuable resource, containing detailed longitudinal data of a large and diverse population.[2,3]

Although the NHS conducts central administrative data collection, data-sharing infrastructure has also evolved through local initiatives, resulting in a patchwork landscape of data extractions without determining what databases or data users exist. This situation has occurred because of three processes. First, following early failures in central information technology programmes,[4] responsibility for technology procurement was delegated to local providers and commissioners. Second, attempts to create national data infrastructure for secondary uses have not achieved public assent, resulting in capability gaps that are increasingly filled by third parties.[5,6] Third, data controller responsibility in NHS England falls to nearly 7000 individual providers who make independent decisions on how data could be used.[7] Overall, decentralisation has allowed procurement to directly support local population needs. However, as discussed in the five year forward view and government reviews, being unable to reach a compromise between over-centralisation and letting a thousand flowers bloom through fragmented local delivery has prevented effective use of unified population data for improving clinical outcomes and reducing health inequalities.[8,9] Inadequate consistency in data controller decision-making processes[10] could also expose patients to risk from privacy breaches, as illustrated by identifiable data exposure to Meta (Facebook) by individual NHS Trusts.[11]

Policy reviews have highlighted privacy and transparency risks in a complex landscape and a need for developing secure population data resources.[1,12,13] At the same time, government strategy aims to secure capabilities such as personalised health intervention, artificial intelligence prediction, and pharmaceutical and life sciences development—all at a population scale.[13] These ambitions share much in common with other countries undergoing digital transformation,[14–16] and are supported by the most extensive package of data infrastructure investment in NHS England history, with up to £200 million announced to support development of secure data environments (SDEs),[17] and a further £480 million for a national federated data platform.[18]

To achieve value, investment must increase data analysis capabilities while striking a balance between privacy and transparency concerns. Policy objectives (panel) should, therefore, be supported by low-level assessment of the current landscape and by assent from an adequately informed public. In this study, we map and characterise all electronic data flows originating from NHS England primary and secondary care providers, flowing to and between visible data consumers. We present three aims: (1) to follow recommendations in the NHS strategy review by Goldacre and Morley[12] for mapping bulk data flows, thus enabling the understanding of privacy risks, capabilities, and positioning of secure data environments; (2) to transparently summarise the complex NHS data landscape; and (3) to build on existing registries of NHS data assets, such as those maintained by Health Data Research UK (HDR UK), but with focus on comprehensiveness, data provenance, and data usage for each asset, through use of systematic mapping techniques. On the basis of our findings, we provide general recommendations to support national data transformation. Finally, we present an interactive public-facing dashboard to visualise data use and to assist with discovery of NHS real-world data assets by the global research community.

Panel: **Summary of relevant UK health data strategy recommendations 2021–22, and relevant questions for landscape mapping, which aim to discover specific details pertinent to strategic recommendations and are used to construct descriptive typology domains**

Public trust in the use of health data

Recommendations

- Improve transparency and encourage patient and citizen engagement[12,13,19,20]
- Move to analytics within controlled, secure data environments[12,13,19,20]
- Reconsider governance models and approach to de-identified data[13,21]

Questions posed

- What data are extracted, who is using it, and for what purposes; how transparent are data extractions and usage?
- Where are secure environments for patient data, and how much data is provisioned securely?
- What control do patients have over consented and non-consented use of de-identified data?

Infrastructural transformation

Recommendations

- Data should be a centralised National Health Service capability, and data flows should be discovered, mapped, and rationalised[12,20]
- New infrastructural solutions to investigate and reduce data and digital inequalities, and avoid digital exclusion[13,20]

Questions posed

- How do content, volume, and distribution of NHS-controlled data flows compare to non-NHS data flows?
- How equitable are existing data extractions by location, extractor, and data content?

Future data-driven capabilities

Recommendations

- Develop multimodal data including genomics to empower researchers and personalised medicine[13,22]
- New guidance and infrastructure to support safe commercial collaboration with life sciences, health technology, and pharmaceutical sectors[12,13,23]
- Support clinical decision makers at every level, and take advantage of artificial intelligence technologies[13,23,24]

Questions posed

- How prevalent are multimodal data linkages?
- How do commercial users access or receive data?
- How does secondary use of data inform clinical care through population health and algorithmic tools?

Methods

Data flows and inclusion and exclusion criteria

We consider the electronic provision of patient-level structured, coded records from NHS England providers for non-direct care uses, which represents data from routine health-care capture but excludes unstructured text records. We term provision of data from one organisation to another as data flow. We include data flows that originate in primary or secondary care providers, which might pass to, and between, subsequent public, academic, non-profit, or commercial entities. Entities might directly procure data from provider health records (ie, data extractor), maintain a standing collection of data for secondary use (ie, database), or use data for a specified purpose (ie, data consumers).[25]

As a snapshot of current infrastructure, we included only systematised data flows or single instance flows between April, 2021, and April, 2022. We excluded entities that collect data by manual collection, as these are not a function of interoperable data infrastructure. We also excluded entities that only provide extraction software, storage, or backup services (eg, cloud providers). Multimodal data, including imaging and genomic data, were considered in the context of linkage to electronic health record data.

See Online for appendix

Information extraction

There are no unified registers of patient data extraction, sublicensing, or usage in NHS England.[12] Figure 1 shows our approach to information discovery. Altogether, we reviewed nine categories of information source, including legal documents produced in respect of the General Data Protection Regulation (GDPR),[26] administrative data use registers, and academic metadata registers. Additional information was requested from 216 secondary care trusts and 106 clinical commissioning groups with freedom of information requests,[27] regarding shared care record data flows and secondary uses. We did scoping reviews of the MEDLINE database to discover named NHS databases, and their subsequent usage in observational research.

We collected characteristics of each data flow, such as data origin and destination, data content and volume, method of data provision or access by consumer, information governance provisions including consent and opt-out mechanisms, how data usage is reported, and how data are used by the destination entity. Information discovery was done between April, 2022, and November, 2022 (JZ, JG, and CO) and is reported in the appendix (p 2).

Reporting, typology, and visualisation

To guide synthesis and narrative reporting of our findings, we summarised themes and recommendations from NHS data strategy publications from 2021 to 2022 (panel). To enable easier description and comparison between data extractors, we created a descriptive typology

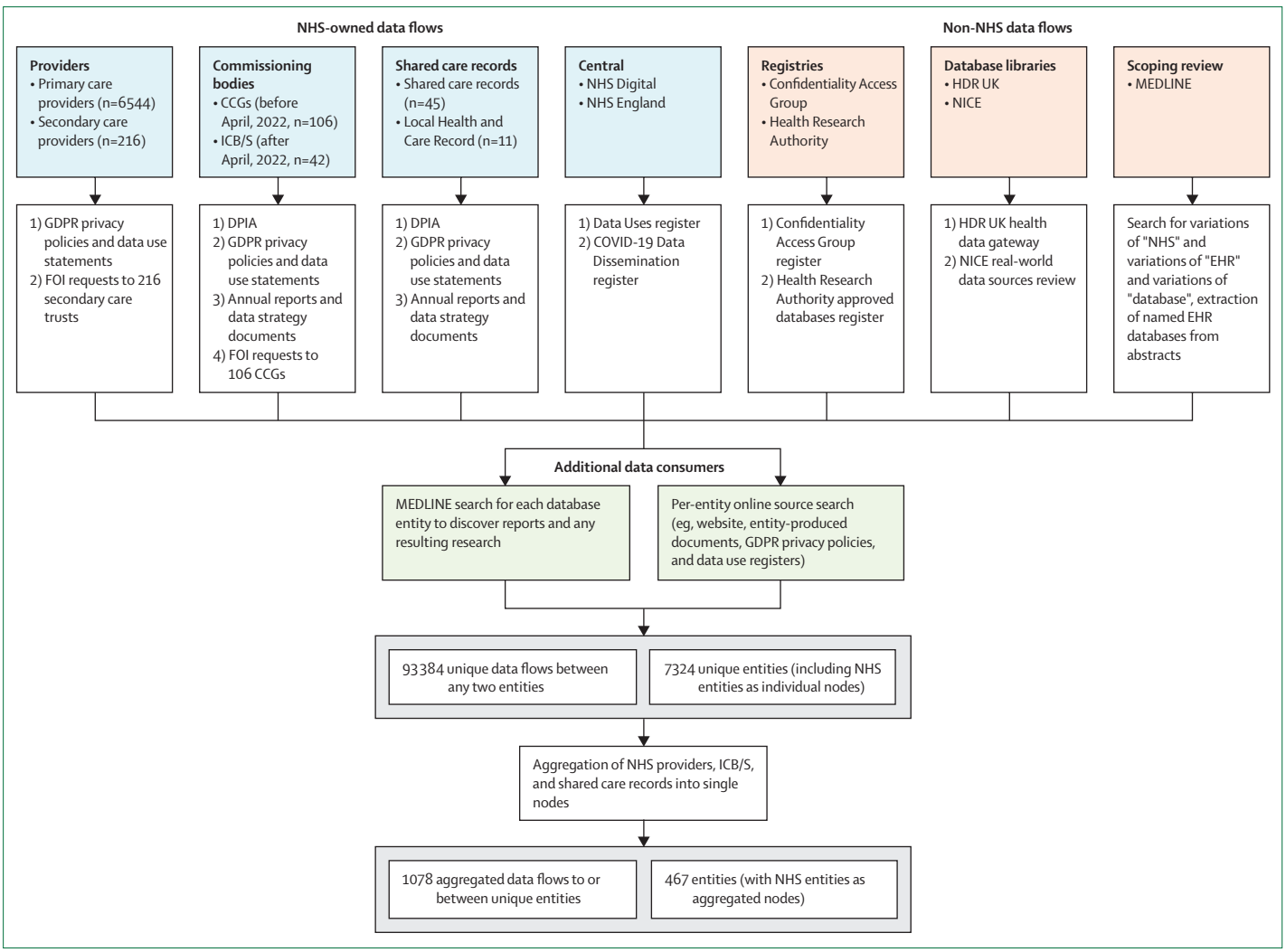

Figure 1: **Flow chart of information sources used in mapping NHS England data flows**
Description of search strategy and sources are found in the appendix (p 2). In online visualisations, NHS entities are represented as aggregated entities to reduce risk of reidentification from within small datasets. CCG=clinical commissioning group. DPIA=data protection impact assessment. EHR=electronic health record. FOI=freedom of information. GDPR=General Data Protection Regulation. HDR UK=Health Data Research UK. ICB/S=integrated care board/system. NHS=National Health Service. NICE=National Institute for Health and Care Excellence.

across important domains.[28] These domains were constructed with relevance to the themes in the panel, while prioritising ease of interpretation by non-experts. Domains include organisation type, data content, data volume and scope, data linkages, method of data provision to consumer, degree of public-facing transparency, model of consent, and onwards consumers and use cases.

We showed discovered information as a graph, with entities as nodes and data flows as relationships. Visualisations (including online dashboards) were created with Python 3.7, Gephi 0.1, and Tableau version 2022.2. To minimise risk of reidentification (ie, connecting small datasets with specific providers), individual care providers and regional bodies are kept anonymous.

Results

Data flows, extractions, and consumers

National data flows are shown in figure 2. Across NHS England, 216 hospital trusts and 6544 primary care providers record health-care interactions for a population of 56 million. All onward data flows originate from four models of data extraction, which are: (1) extraction of structured clinical codes from primary care electronic health records;[7] (2) administrative data collection by NHS Digital from secondary care, including main diagnoses for individual care episodes;[29] (3) data aggregated within regional shared care record data warehouses, representing capture of standardised messages from primary and secondary care electronic health records;[30] and (4) proprietary secondary care data pipelines, generally extracting data of higher temporal and

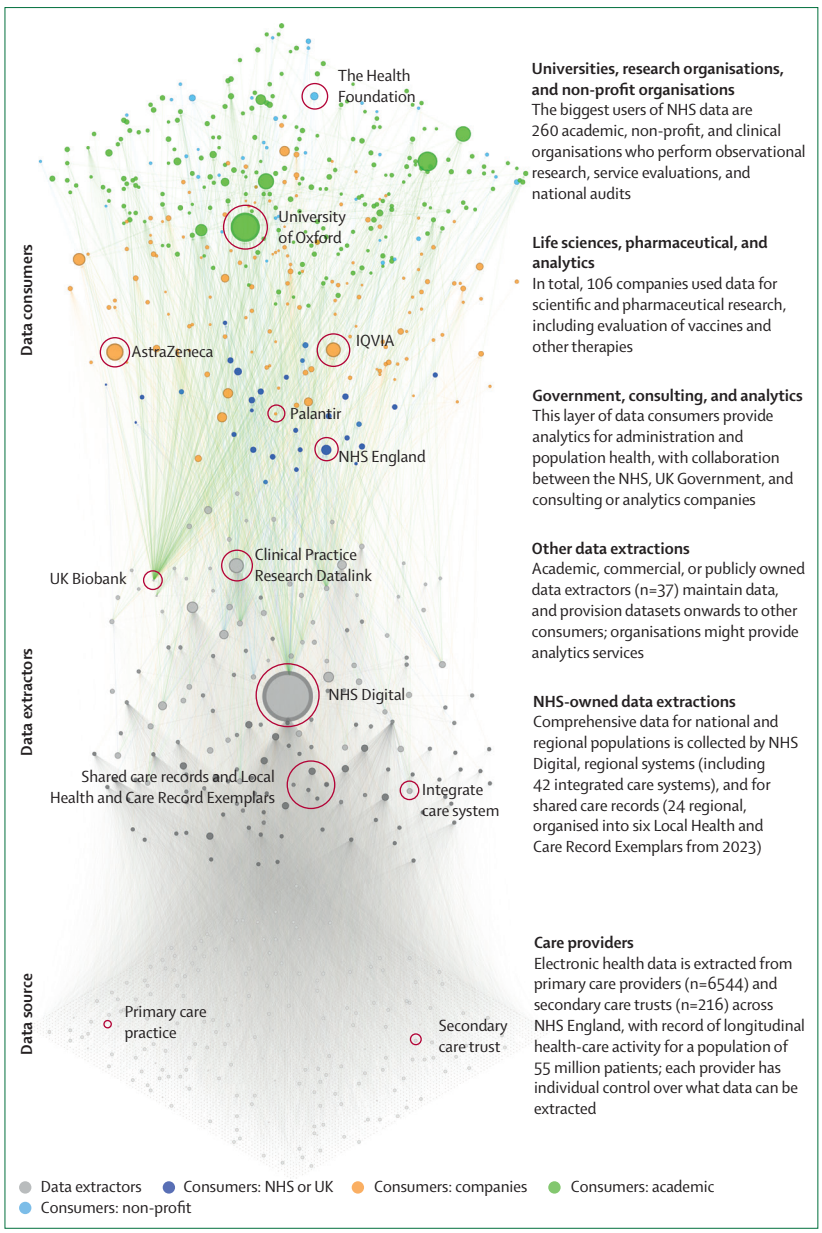

Figure 2: **Electronic patient data flows in NHS England**
Data flows go upwards and are coloured by destination. For data source and extractors, node size is proportional to population catchment (eg, NHS Digital=55 million). For data consumers, node size is proportional to the number of projects (eg, University of Oxford=178). NHS=National Health Service.

For more on the **Health Informatics Collaborative** see https://hic.nihr.ac.uk/

For more on the **UK Biobank** see see https://www.ukbiobank.ac.uk/

For more on **Genomics England** see https://www.genomicsengland.co.uk/

information resolution[31,32] when compared with administrative datasets.

Extracted data feed a vast ecosystem of secondary uses, which include at least 460 non-NHS organisations who have accessed, maintained, or used NHS data since April, 2021. At the far end of the data flow chains, consumers include researchers from 216 universities or academic organisations; 143 pharmaceutical, life sciences, data analytics, and consulting companies; and 44 non-profit organisations. Figure 3 shows the top consumers and main use cases in each category.

More than 95% of consumers collect these data indirectly via data extractor intermediaries (eg, NHS Digital, regional NHS bodies, and 37 non-NHS organisations). Although the median data flow chain consists of three entities (ie, provider, extractor, and consumer), we discovered 56 (12%) of 460 consumers sharing data with at least one further consumer.

Types of data extractor

Data extractors are key intermediary nodes that maintain and provide datasets to consumers. We describe eight distinct extractor types (figure 4), and individual extractors are described in the appendix (p 5).

NHS Digital hosts the only whole-population secondary care datasets, derived from administrative collections, and maintains the General Practice Extraction Service Data for Pandemic Planning and Research (GDPPR) from primary care extractions for COVID-19 use. Other data of national scope is held by primary care research databases, which extract from differing numbers of practices across the country, with the largest, the Clinical Practice Research Datalink, supporting 18 million active patients.

12 commercial data extractors can act as brokers (ie, licensing datasets to consumers), including databases run by companies such as IQVIA[33] and cegedim,[34] but can also provide specific services to customers. Agreements are maintained with individual providers.

We found 24 active shared care records systems that hold data for direct-care purposes. System suppliers might offer additional population health analytics capabilities. By the end of 2023, many systems will be centralised into local health and care record regions, with catchments of up to 10 million patients. A subset of these systems enable access to hosted data for research purposes.

The NHS is administered through 42 integrated care systems that use linked data for commissioning and population health uses, supported by commissioning support units and analytics companies. These data extractions have population coverage for each geographical region. Some datasets are also made available for academic users. Smaller volume secondary care data pipelines support seven academic research collaboratives (eg, the Health Informatics Collaborative), which curate cohort data on the basis of thematic inclusion criteria, and 12 secondary care centre databases accessible to research users. Finally, two prospective cohorts (ie, the UK Biobank and Genomics England) perform linkage of genomics data to primary care and NHS Digital secondary care data to enrich cohort follow-up.

Balance of data assets and distribution

Data extractors differ by type and volume of maintained data, and act as flow multipliers, by each enabling multiple distribution routes. This results in differing availability and usage of different data types (figure 5). The most prevalent maintained data are from primary

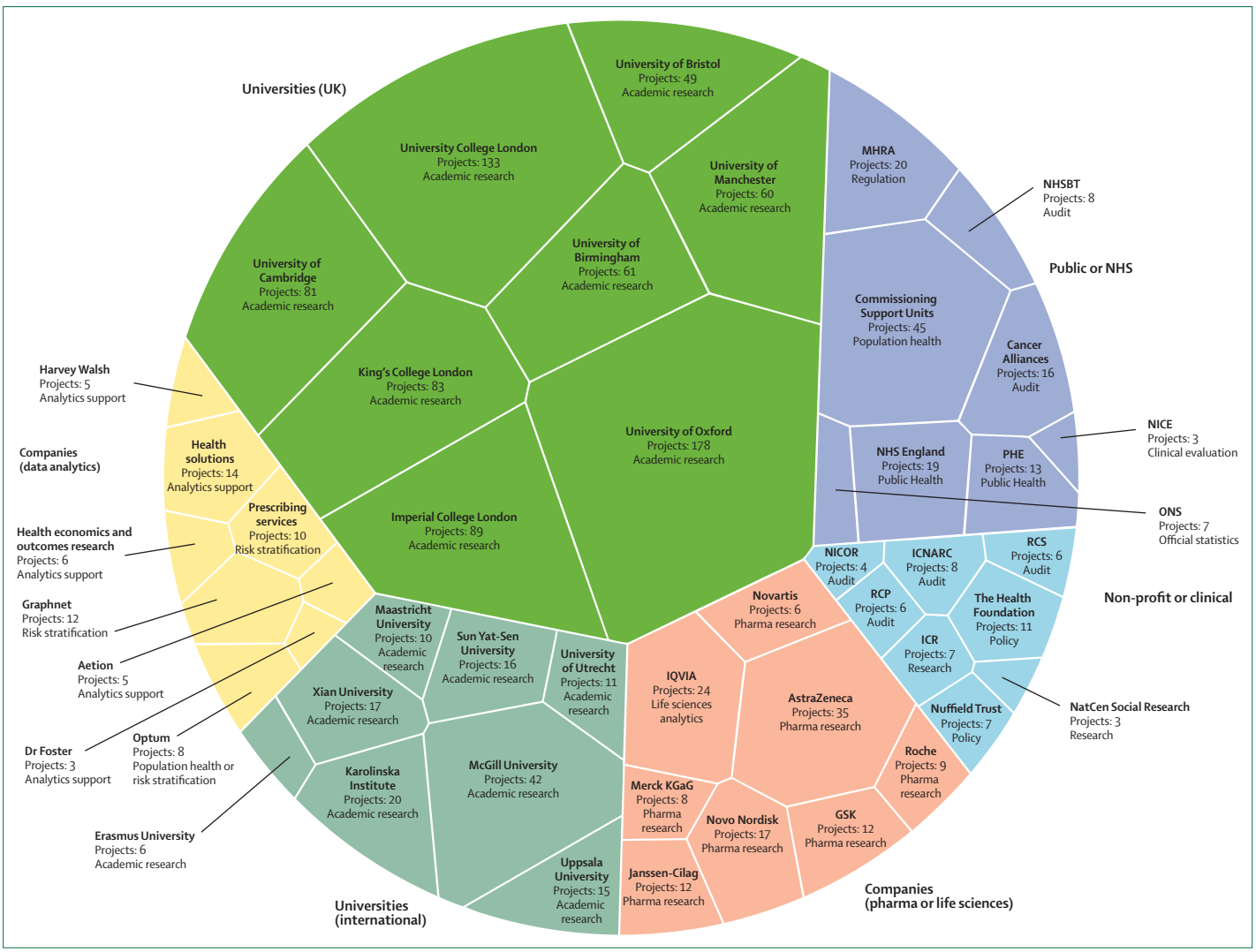

Figure 3: **Voronoi chart showing eight top consumers for NHS data across each of six categories, by number of discovered projects during the study period**
Description shows dominant form of data use. Projects for data consumers might include any research study, research publication, audit, listed operational or public health analyses, and ongoing public, academic, or commercial partnerships. Projects are derived from database websites, data registers, and a scoping literature search (appendix pp 2–4). Research involvement defined from lead or applicant organisation, first or senior author in academic publications, or named involvement in data analysis, and does not directly represent funding source for projects. Data usage might be under-represented due to reporting limitations. GSK=GlaxoSmithKline. ICNARC=Intensive Care National Audit and Research Centre. ICR=Institute of Cancer Research. MHRA=Medicines and Healthcare Products Regulatory Authority. NHS=National Health Service. NICE=National Institute for Health and Care Excellence. NICOR=National Institute for Cardiovascular Outcomes Research. ONS=Office for National Statistics. PHE=Public Health England. RCP=Royal College of Physicians. RCS=Royal College of Surgeons.

care. Whole-population primary care data are available only for COVID-19 research via GDPPR, or the federated analytics platform OpenSAFELY. Other data, held in commercial and academic databases of national scope (n=7), are extracted for a cumulative, overlapping, active population of 76 million patients, but with a median independent database size of 13 million (IQR 11–15 million) active patients. Of data extractions that were reported by primary care practices, 28% report extraction by two databases and 21% reported data extractions by three or more, suggesting substantial duplication between databases.

Administrative secondary care data are the only general use whole-population data asset that is held by NHS Digital. Partial copies are permanently maintained in at least 12 additional research databases and in regional care systems. Overall, primary care and administrative secondary care data were distributed to 90% of unique consumers in the study period.

For more on **OpenSAFELY** see https://www.opensafely.org/

Conversely, we estimate the median extraction size from hospital data pipelines to be less than half a million patient episodes. Other more granular secondary care data are found in shared care record data warehouses, but only four shared care record databases support secondary use for research, including two for COVID-19 usage only (total 3·5 million patients).

Linkage to multimodal imaging or genomics data is found exclusively at prospective cohorts, local centres,

Data extractors	Size	Scope			Linkage			Provision			Ownership			Reporting			Consent		Consumers				
	Median	National	Regional	Local	NHS Digital	Imaging	Genomic	Majority SDE	Mixed	Majority non-SDE	Public, NHS, government	Academic or university	Commercial	Approvals or uses register	Publications or outcomes	Non-specific statement	Consent	Opt-out	Local academic	Collaborative academic	Local public	Government	Commercial
NHS Digital	Whole population	✓								✓	✓			✓				✓	✓	✓	✓	✓	✓
OpenSAFELY*	Whole population	✓			✓			✓			✓	✓		✓				✓		✓		✓	
Primary care research	12 million	✓			✓				✓		✓	✓		✓				✓		✓		✓	✓
Commercial data brokers	5·5 million	✓			✓		✓			✓			✓			✓		✓		✓				✓
Shared care record systems	3·5 million		✓						✓		✓		✓		✓			✓	✓		✓		✓
Regional systems	2 million		✓		✓					✓	✓				✓			✓	✓		✓		
Themed research cohorts	<500 000	✓	✓					✓				✓			✓			✓		✓			
Secondary care centres	<500 000			✓		✓		✓				✓			✓			✓	✓				
Prospective cohorts	<500 000	✓			✓	✓	✓	✓			✓	✓	✓	✓			✓			✓			✓

Data extraction model

Primary care extractions; Secondary care extractions; Shared care records; Administrative secondary care

GDPPR (COVID-19 only)

OpenSAFELY federated analytics platform (COVID-19 only)

Figure 4: **Data extractor typology showing eight distinct types**

Extractors show unique characteristics across multiple domains. GDPPR=General Practice Extraction Service Data for Pandemic Planning and Research. NHS=National Health Service. SDE=secure data environment. *OpenSAFELY is shown for comparison only; rather than extracting data, it enables federated analytics across primary care vendor databases.

and one commercial data extractor (23andMe).[35] The largest multimodal cohort includes half a million patients in the UK Biobank, linking genomic data to primary and administrative secondary care data, and is the single most influential data distributor, supplying 190 different consumers.

Consent and reporting transparency

Within discovered extractors, only two prospective cohort databases and one commercial genomics database extract and distribute patient data with explicit consent, accounting for less than 0·5% of maintained data assets. NHS Digital facilitates consented linkage to external research cohorts. Most extractions occur under alternative legal provisions for performing tasks in the public interest.[36] Patient control over data use relies on opt-out mechanisms, at the levels of primary care extraction, primary care provision to shared care records, and through a central record held on the NHS spine.[37]

For more on **DataInsights** see https://DataInsights.uk

Care providers are expected to report possible uses of patient data across primary care practice websites. Reporting of data extractions through primary care practice websites (n=6544) was estimated at 63% of what would be expected from reports of practice enrolment by databases. Secondary care providers report potential for data to be used in research, but with no specificity to projects or consumers. For data extractors, NHS Digital, primary care research databases, and prospective cohort studies provide public facing registers of active projects. Most commercial data extractors provide only non-specific description of onwards data usage.

For more on the **HDR UK gateway** see https://www.healthdatagateway.org/

Secure data environments

Although dedicated research platforms with secure access to patient data are traditionally known as trusted research environments, the NHS now considers all privacy-focused data analysis environments under the term SDE.[38] The greatest volume of linked data can be accessed in an NHS Digital internal environment, including whole-population data for COVID-19 analyses. However, 102 (78%) of 130 NHS Digital data consumers, including 31 (89%) of 35 companies, opted for data to be transferred outside of an SDE. We discovered 20 additional environments (figure 5) that otherwise fulfilled SDE criteria, accounting for data provision to 35% of unique consumers.

Public-facing dashboard

We present interactive visualisations online on the DataInsights website. The website is structured across three infographics written for non-experts, including an explainer of different data types, flows, provenance, and destinations; a comprehensive description of systematically discovered electronic health record databases that are accessible to external researchers; and a cross-section of the largest users of NHS data. Where included, more detailed metadata, including covariable information, can be discovered through the HDR UK gateway.

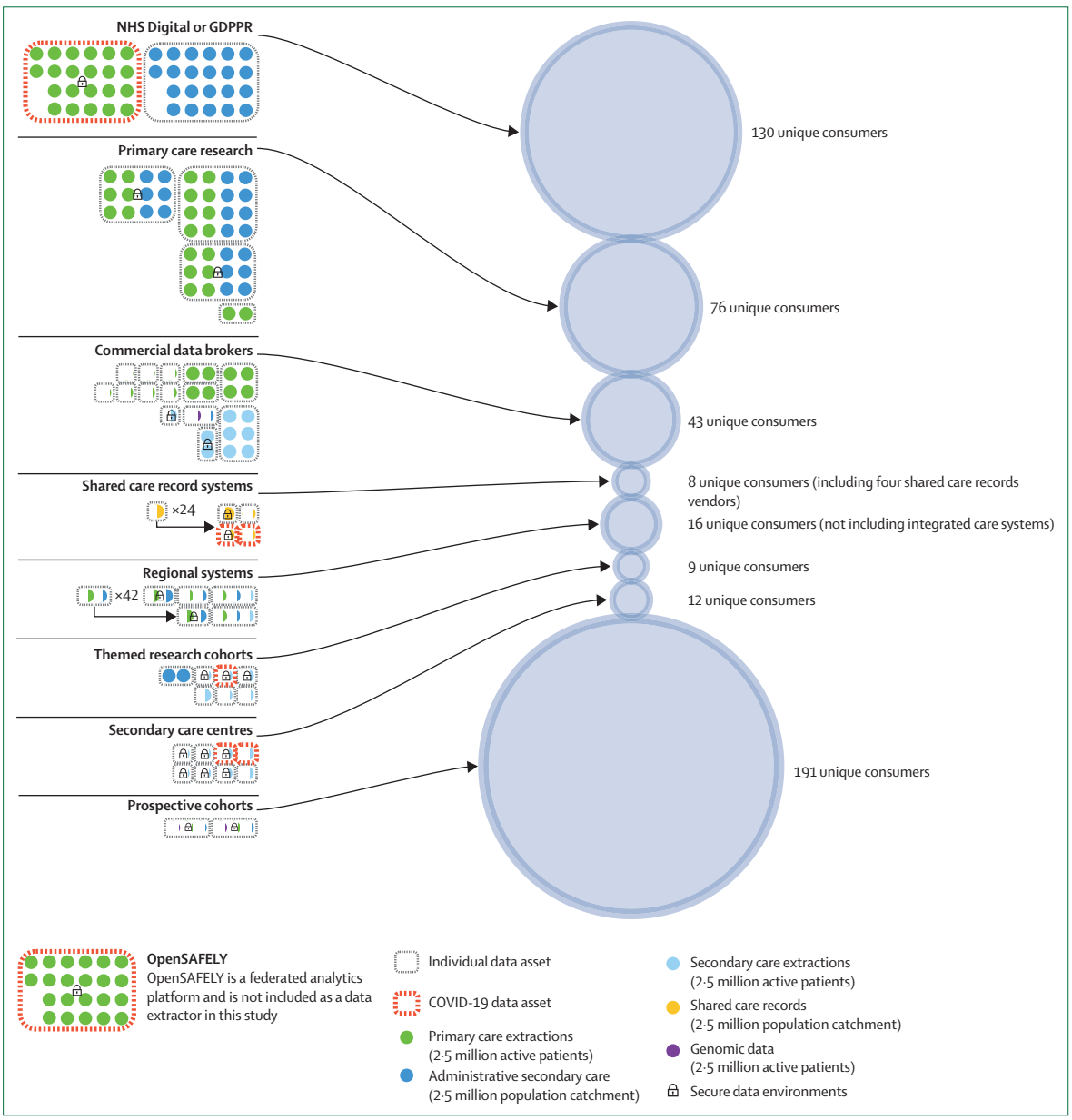

Figure 5: **Individual data assets per extractor type, showing volume of data types and linkages**
Relative consumption shown as number of unique data consumers. Few shared care record and regional systems host their datasets as research environments for external users (shown as separate assets). GDPPR=General Practice Extraction Service Data for Pandemic Planning and Research.

Risks to public trust in data use

We have described a complex landscape that contains hundreds of organisations positioned along multistage data flow chains. The use of de-identified data without explicit consent is a point of controversy within post-pandemic[6] and historical[5,39] failures of NHS data programmes. The current landscape shows failures in transparency and privacy that risk compromising public trust.

Data usage most often occurs two or three interactions down a chain. At each stage, data flows have a one-to-many relationship. For any patient in NHS England, data flows to a minimum of two and up to 16 potential data extractors, each with their own ecosystem of subsequent data flows. These stages of multiplicative data distribution place patients at considerable distance from data usage. Furthermore, we found reporting of data uses to be incomplete or having low specificity, including boilerplate notices that state data are used for research, which risks violation of the no surprises principle within data protection legislation,[37,40] and places the onus on patients to actively investigate how their data is being used.

The majority of the UK public support the use of de-identified data for public benefit or to advance medical knowledge but are more cautious about use of data for commercial profit, reflecting a similar stance to populations worldwide.[41–44] Public research and dialogue, including that commissioned by the National Data Guardian, find assent to be predicated on full transparency, requiring clear distinction between specific use cases.[19,42,45,46] Data ultimately reflect individuals, and when transparency is low, patients are unable to understand what inferences might be drawn from their digital data, thus undermining autonomy and trust.

Public assent is important in the context of de-identified data. Data flows in this study are either anonymised through personal identifiable data removal or pseudonymised with keys for reidentification or linkage. In either case, reidentification through temporal characteristics of events or isolated rare conditions are a recognised risk.[47,48] Clarity over risk mitigation is especially important if patients have little control or statutory protection over how data are used. The Goldacre review establishes potential risks in unaudited bulk data flows and produces strong recommendations for restructuring data into a small number of secure environments.[38] We found physical data transfers outside of SDEs to be the majority occurrence. In data flows and usage that are audited, previous investigative research has uncovered numerous breaches of data contracts and confidentiality agreements, including in 33 (100%) audited organisations who used NHS Digital data over the same period as this study.[49,50] Breaches are also likely to occur if data flows are unaudited, which is a majority of the landscape. The possibility of unobserved data breaches risks additional damage to patient trust. In addition, although the risk of patient reidentification by a malicious individual or group is low, this risk is magnified if numerous data breaches are occurring across hundreds of data flow chains.

In consideration of persistent risks, the Goldacre review further recommends a category of de-identified but re-identifiable data.[12] New guidance from the Information Commissioner's Office proposes factors for testing risk of reidentification,[21] but these conditions are open to interpretation. In the USA, the Health Insurance Portability and Accountability Act, like the GDPR, does not apply to de-identified data; however, there is more structured focus on delineating technical de-identification best practices, alongside expert peer review to explore reidentification risks,[51,52] which could produce greater uniformity in practice and increase confidence in anonymisation.

Robust opt-out mechanisms can help to maintain trust in de-identified data use. However, positioning of opt-outs has three potential risks. First, although there are clear differences in public assent for different uses of data, these uses are not considered by blanket opt-outs positioned at extraction level. Second, data flow to shared care records might be controlled by an opt-out of information exchange for direct care, potentially asking patients to choose between having data shared for both clinical care and secondary uses, or not having data shared at all. Third, even if a patient opts out at all levels, de-identified data could still flow to numerous secondary uses.[53,54]

Data volume hides insufficient diversity in information and population

Our findings lend broad support to expansion of SDEs. However, investment must consider necessity for additional nodes of data aggregation, linkage, and provision. At face value, the NHS possesses enormous data resources, but these resources partly reflect duplication, rather than information or population diversity.

Enormous quantities of primary care data are segmented across numerous databases. Similarly, databases hold partial, duplicated NHS Digital datasets in different locations for onward provision. Present NHS Digital infrastructure and the federated OpenSAFELY platform are technologically capable of supporting secure provision of whole-population linked data as a general research asset (ie, a capability shared by only a handful of countries with much smaller populations),[55] but only for COVID-19 uses. Limitations in national capability, therefore, reflect public concerns regarding risk,[6] rather than availability of data infrastructure. Conversely, we find technological gaps in access to secondary care electronic health record data, in which multiplicity of vendor systems makes interoperability a continued challenge,[56,57] with high barriers of entry limiting success to a few digitally mature centres.

This imbalanced landscape has implications for effective and equitable data use. Insufficient information diversity affects research capabilities, as individual data sources are known to suffer from quality issues and missing data.[58–63] Complementary linked data types enable complete capture of patient lifetime journeys and ensure that NHS services meet the needs of everyone in the population. In particular, the need for high-quality secondary care data was exemplified during the COVID-19 pandemic, when in-hospital trajectories were crucial for informing research and planning.[63]

Reduced population diversity risks negative bias resulting from differences between high and low data flow density areas. Data flows are determined by local data-sharing practices, by presence of digitally mature academic centres, and recruiting practices for cohort studies.[64,65] These upstream factors are known to result in unrepresentative data that adversely affect research, pharmaceutical evaluation, and artificial intelligence development.[66,67] Routinely collected NHS data have particular value due to universal health-care access, especially compared with insurance-based systems, in which data aggregation largely represents well-served populations.[68,69] These priorities are reflected in national strategy aiming to reduce data and digital disparities.

Overall, new SDEs will enable further nodes of access to data that are already widely available but might not

widen information or population diversity. Work is required in data extraction technologies and in expanding multimodal resources that support personalised medicine interventions. One possible route for improving secondary care electronic health record data availability is through existing shared care records, which have developed separately from research infrastructure. However, legal and governance provisions for shared care records might not consider data consumption for secondary uses. For multimodal data, high profile genomics projects such as Our Future Health and the expansion of radiomics programmes, such as the National COVID-19 Chest Imaging Database,[70] are key to addressing data imbalance. For now, the best way to use available resources might be through patient and public programmes to achieve assent for expanding available SDEs that already contain linked whole-population data.

Value implications across data flow chains

Data flow mapping allows examination of value gain and loss across each chain. Although a quantitative analysis is outside the scope of this study, several findings warrant additional discussion. Data flow chains carry substantial monetary value, which are multiplicative at each stage through sublicensing fees and commercial consumption. For the largest databases, costs are ultimately borne by researchers or companies that wish to access data, which is described as prohibitive for many in academia[71] (eg, multi-study licensing fees between £75 000 and £330 000).[72] These costs cover infrastructure and administration, but could also produce net income, particularly for commercial brokers. For consumers, data access might additionally support revenue-generating services.

By contrast, value return to patients, care providers, and the NHS is a minority proportion of this landscape.[73] Some databases offer financial incentive packages for providers (eg, The Health Improvement Network offering either £600 or three iPads for data from 10 000 to 15 000 patients), but these exchanges do not scale to propagation of revenue-generating interactions with consumers.

When considering value for patient care and population health, most data are used for observational research, with public benefits across a long time horizon and real-world impacts that are difficult to quantify. The direct data-driven interventions that are the focus of NHS strategy are seen only in a small number of suppliers of population health and risk stratification services or at the level of regional commissioners. Although a focus on analytics environments is optimal for performing observational research, greater value might be generated through platform infrastructure. Platform infrastructure refers to components that support engineering and maintenance of continuous data flows and tools for entire project lifecycles, including an implementation and delivery stage, beyond that supported by SDE-hosted research.[74–76]

Finally, we consider value loss. Each database node in a chain requires a substantial monetary cost, which is borne by academic or public funding for research aims. Arguably, a new database that duplicates data already found elsewhere adds little value. Moving research questions into existing SDEs might reduce spending on new and costly data infrastructure, improve collaboration and reproducibility, and reduce the need for bulk physical data transfers. These findings support recommendations in the Goldacre review for avoiding new bulk data aggregations, instead restructuring existing data flows and analyses into a small number of SDEs with advanced capabilities.[12]

Recommendations for data transformation

We have translated our findings into six recommendations pertinent to the NHS and national data initiatives in general. First, public transparency should not require investigative discovery. With numerous nodes of dissemination, distal data uses should be adequately reported to ensure transparency and safeguard against data breaches. Future work should focus on the extent and methods of reporting.

Second, opt-out conditions should be set at the level of distribution to types of consumers, rather than at extraction. As capabilities advance, binary extraction opt-outs will limit patient autonomy and restrict access to key data-driven interventions.

Third, improving and widening usage of existing infrastructure is a priority. In the NHS, ongoing programmes of public outreach and education could enable availability of linked assets in NHS Digital and OpenSAFELY for general uses. Precedent for mandating usage of such environments should be created before establishing new SDEs.

Fourth, new data infrastructure must focus on expanding capabilities for extracting untapped secondary care electronic health record data, and increasing multimodal data availability, rather than reshuffling existing assets into additional nodes of dissemination. This expansion might require technologically individualised solutions across regions. Most pertinently, a new national federated data platform might enable analytics across regional data environments, but whether the resulting data will differ from that already held by the NHS is unclear. Federation of data will bring advantages for data privacy, while reducing the need for bulk data transfers, but will also rely on regional participation and increase local infrastructure complexity. Positioning of a new federated data platform is shown in the appendix (p 10).

Fifth, to increase value return to patients and providers, infrastructure should focus on intervention, rather than on analysis capabilities. Required capabilities include increasing data provision cycle time, regulatory and governance optimisation for product lifecycles, and introducing production capabilities for artificial intelligence.[76–78] Regional centres that already extract data for population health analytics are well placed to develop such infrastructure. This is a current investment focus in the NHS through subnational SDEs.[79]

For more on **Our Future Health** see https://ourfuturehealth.org.uk/

For more on **The Health Improvement Network** see https://www.the-health-improvement-network.com/gp

Finally, monetary value transfer across the entire data flow landscape should be assessed to quantify value returned to the health-care system. This assessment is essential in a health system that is struggling financially and might not receive return on investment if most value is generated two or three stages into a data flow chain. This assessment is also a necessary basis from which to institute beneficial models of revenue return to patients and providers in the NHS.[73]

Strengths and limitations

Our consideration of data flows enables unique understanding of national data provenance and utilisation. Systematic mapping of the data landscape increases transparency and provides objective understanding for policy makers. Our approach is primarily limited by possibilities for information discovery. We are only able to present a macro view of the landscape and might lose smaller, local data transfers. There is potential for under-reporting at all stages, particularly in commercial data flows. We do not include consideration of other UK nations due to reporting differences. We do not include a large amount of routine health data available in social care through unlinked imaging datasets, patient reported data, or manually curated datasets. Similarly, our description of data uses is broadly classified (figure 3), and detailed analysis of how individual consumers use patient data is outside the scope of this study. Finally, SDEs might be under-reported, as we rely on self-reporting against current definitions.

Conclusion

Public reaction to proposed NHS-led data projects suggests an uncomfortable possibility that the extent and methods of patient data dissemination shown in this study far exceed present awareness. Instead, we argue that a process of restructuring is required to ensure security, diversity, and return of value to patients and providers. Administrative regions with responsibility for commissioning are an important node for investment due to existing data flows, public support for population health uses, and proximity to the clinical front line for delivering actionable insights. In general, public spending must deliver more than duplicative analytics infrastructure. Bottlenecks exist in the use of existing infrastructure, public assent, data extraction technologies, multimodality, and models of value return to the NHS. Investment into data transformation must focus on these foundational components.

Contributors

JZ contributed towards the conception of the manuscript. JZ and JM contributed towards the methodology. JZ, JG, and CO contributed towards the data collection and analysis. All authors contributed towards the writing and approval of drafts and revisions. The final manuscript was approved by all authors.

Declaration of interests

HA is Chief Scientific Officer of Preemptive Health and Medicine and Flagship Pioneering. JM was paid directly for giving a lecture at Health Education England on the topic of artificial intelligence in the NHS. JM has been a member of the INSIGHT DataTAB for HDR UK. This paper references the Goldacre Review, for which JM was a coauthor. All other authors declare no competing interests.

Acknowledgments

JZ acknowledges funding from the Wellcome Trust (203928/Z/16/Z) and support from the National Institute for Health Research Biomedical Research Centre based at Imperial College NHS Trust and Imperial College London. JM is a Wellcome Trust Doctoral Fellow.

References

1 Department of Health and Social Care. A plan for digital health and social care. June 29, 2022. https://www.gov.uk/government/publications/a-plan-for-digital-health-and-social-care/a-plan-for-digital-health-and-social-care (accessed June 7, 2023).
2 Bagenal J, Naylor A. Harnessing the value of NHS patient data. *Lancet* 2018; **392:** 2420–22.
3 Fontana G, Ghafur S, Torne L, Goodman J, Darzi A. Ensuring that the NHS realises fair financial value from its data. *Lancet Digit Health* 2020; **2:** e10–12.
4 Justinia T. The UK's national programme for IT: why was it dismantled? *Health Serv Manage Res* 2017; **30:** 2–9.
5 Godlee F. What can we salvage from care data? *BMJ* 2016; **354:** i3907.
6 Burki T. Concerns over England's new system for collecting general practitioner data. *Lancet Digit Health* 2021; **3:** e469–70.
7 Bradley SH, Lawrence NR, Carder P. Using primary care data for health research in England—an overview. *Future Healthc J* 2018; **5:** 207–12.
8 NHS England. Five year forward view. October, 2014. https://www.england.nhs.uk/wp-content/uploads/2014/10/5yfv-web.pdf (accessed June 7, 2023).
9 Parliament UK. The long-term sustainability of the NHS and adult social care: chapter 5: innovation, technology and productivity. April, 2017. https://publications.parliament.uk/pa/ld201617/ldselect/ldnhssus/151/15108.htm ((accessed June 7, 2023).
10 Information Commissioner's Office. Contractual liability in data sharing agreements. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/contractual-liability-in-data-sharing-agreements/ (accessed May 1, 2023).
11 Das S. NHS data breach: trusts shared patient details with Facebook without consent. May 27, 2023. The Guardian. https://www.theguardian.com/society/2023/may/27/nhs-data-breach-trusts-shared-patient-details-with-facebook-meta-without-consent (accessed June 7, 2023).
12 Goldacre B, Morley J, Hamilton N. Better, broader, safer: using health data for research and analysis. April, 2022. Department of Health and Social Care. https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis/better-broader-safer-using-health-data-for-research-and-analysis (accessed June 7, 2023).
13 Department of Health and Social Care. Data saves lives: reshaping health and social care with data. June 15, 2022. https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data (accessed July 7, 2022).
14 Directorate General for Parliamentary Research Services. EU health data centre and a common data strategy for public health. Brussels: European Union, 2021.
15 Office of the Chief Technology Officer. Leveraging data for the nation's health. December, 2019. US Department of Health and Human Services. https://www.hhs.gov/sites/default/files/master-future-state-508.pdf (accessed June 7, 2023).
16 Pacific Health Information Network. Evaluation and Renewed Vision and Strategy for the Pacific Health Information Network (PHIN). December, 2018. https://www.who.int/docs/default-source/wpro---documents/dps/evaluation-and-renewed-vision-and-strategy-(2019-2021)-for-the-pacific-health-information-network-(phin).pdf?sfvrsn=c48bf1f7_2 (accessed June 7, 2023).
17 Department of Health and Social Care. £260 million to boost healthcare research and manufacturing. March 2, 2022. https://www.gov.uk/government/news/260-million-to-boost-healthcare-research-and-manufacturing (accessed June 7, 2023).
18 NHS England. NHS federated data platform and associated services. Jan 10, 2023. https://www.find-tender.service.gov.uk/Notice/000669-2023 (accessed June 7, 2023).

19 Office of the National Data Guardian for Health and Social Care. National Data Guardian 2021–22 report. Aug 30, 2022. https://www.gov.uk/government/publications/national-data-guardian-2021-2022-report (accessed June 7, 2023).

20 Department of Health and Social Care. Putting data, digital and tech at the heart of transforming the NHS. Nov 23, 2021. 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 (accessed June 7, 2023).

21 Information Commissioner's Office. Privacy-enhancing technologies (PET): anonymisation, pseudonymisation, and privacy enhancing technologies guidance (draft). September, 2022. https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf (accessed June 7, 2023).

22 NHS England. Accelerating genomic medicine in the NHS. Oct 31, 2022. https://www.england.nhs.uk/long-read/accelerating-genomic-medicine-in-the-nhs/#foreword (accessed June 7, 2023).

23 Department for Digital, Culture, Media and Sport. UK digital strategy. Oct 4, 2022. https://www.gov.uk/government/publications/uks-digital-strategy/uk-digital-strategy (accessed June 7, 2023).

24 Central Digital and Data Office. Roadmap for digital and data, 2022–25. June 9, 2022. https://www.gov.uk/government/publications/roadmap-for-digital-and-data-2022-to-2025 (accessed June 7, 2023).

25 Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. *J Manage Inf Syst* 1996; **12:** 5–33.

26 Rumbold JMM, Pierscionek B. The effect of the General Data Protection Regulation on medical research. *J Med Internet Res* 2017; **19:** e47.

27 Savage A, Hyde R. Using freedom of information requests to facilitate research. *Int J Soc Res Methodol* 2014; **17:** 303–17.

28 Collier D, Laporte J, Seawright J. Typologies: forming concepts and creating categorical variables. In: Box-Steffensmeier JM, Brady HE, Collier D, eds. The Oxford handbook of political methodology, 1st edn. Oxford: Oxford University Press, 2009: 152–73.

29 NHS Digital. NHS Digital: secondary use services. February, 2022. https://digital.nhs.uk/services/secondary-uses-service-sus (accessed June 7, 2023).

30 NHS England. Interoperability. https://www.england.nhs.uk/digitaltechnology/connecteddigitalsystems/interoperability/ (accessed Sept 5, 2023).

31 Smith DA, Wang T, Freeman O, et al. National Institute for Health Research Health Informatics Collaborative: development of a pipeline to collate electronic clinical data for viral hepatitis research. *BMJ Health Care Inform* 2020; **27:** e100145.

32 Harris S, Shi S, Brealey D, et al. Critical Care Health Informatics Collaborative (CCHIC): data, tools and methods for reproducible research: a multi-centre UK intensive care database. *Int J Med Inform* 2018; **112:** 82–89.

33 Myland M, O'Leary C, Bafadhal B, et al. IQVIA Medical Research Data (IMRD). In: Sturkenboom M, Schink T, eds. Databases for pharmacoepidemiological research, 1st edn. Cham: Springer Nature Switzerland, 2021: 67–76.

34 cegedim. THIN: The Health Improvement Network. https://www.cegedim-health-data.com/cegedim-health-data/thin-the-health-improvement-network/ (accessed Sept 5, 2023).

35 23andMe. Research. https://www.23andme.com/en-gb/research/ (accessed Sept 5, 2023).

36 Health Research Authority. Legal basis for processing data. May 8, 2018. https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/data-protection-and-information-governance/gdpr-detailed-guidance/legal-basis-processing-data/ (accessed June 7, 2023).

37 Evans H. Using data in the NHS: the implications of the opt-out and GDPR. May 25, 2018. https://www.kingsfund.org.uk/publications/using-data-nhs-gdpr (accessed July 7, 2022).

38 Department of Health and Social Care. Secure data environment for NHS health and social care data—policy guidelines. Dec 23, 2022. https://www.gov.uk/government/publications/secure-data-environment-policy-guidelines/secure-data-environment-for-nhs-health-and-social-care-data-policy-guidelines#:~:text=Secure%20data%20environments%20are%20data,the%20data%20leaving%20the%20environment (accessed June 7, 2023).

39 BBC News. Google DeepMind NHS app test broke UK privacy law. July 3, 2017. https://www.bbc.co.uk/news/technology-40483202 (accessed Sept 5, 2023).

40 NHS Digital. Protecting patient data. Sept 6, 2022. https://digital.nhs.uk/services/national-data-opt-out/understanding-the-national-data-opt-out/protecting-patient-data (accessed Sept 8, 2022).

41 Atkin C, Crosby B, Dunn K, et al. Perceptions of anonymised data use and awareness of the NHS data opt-out amongst patients, carers and healthcare staff. *Res Involv Engagem* 2021; **7:** 40.

42 Jones LA, Nelder JR, Fryer JM, et al. Public opinion on sharing data from health services for clinical and research purposes without explicit consent: an anonymous online survey in the UK. *BMJ Open* 2022; **12:** e057579.

43 Richter G, Borzikowsky C, Lieb W, Schreiber S, Krawczak M, Buyx A. Patient views on research use of clinical data without consent: legal, but also acceptable? *Eur J Hum Genet* 2019; **27:** 841–47.

44 Trinidad MG, Platt J, Kardia SLR. The public's comfort with sharing health data with third-party commercial companies. *Humanit Soc Sci Commun* 2020; **7:** 149.

45 van Staa T-P, Goldacre B, Buchan I, Smeeth L. Big health data: the need to earn public trust. *BMJ* 2016; **354:** i3636.

46 Hopkins Van Mil. Putting good into practice—a public dialogue on making public benefit assessments when using health and care data. April, 2021. https://www.gov.uk/government/publications/putting-good-into-practice-a-public-dialogue-on-making-public-benefit-assessments-when-using-health-and-care-data (accessed June 7, 2023).

47 Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc* 2010; **17:** 169–77.

48 El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. *PLoS One* 2011; **6:** e28071.

49 Oxford E. Hundreds of patient data breaches are left unpunished. *BMJ* 2022; **377:** o1126.

50 Banner N. NHS data breaches: a further erosion of trust. *BMJ* 2022; **377:** o1187.

51 Office for Civil Rights. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule. Oct 25, 2022. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (accessed June 7, 2023).

52 Mandl KD, Perakslis ED. HIPAA and the leak of "deidentified" EHR data. *N Engl J Med* 2021; **385:** e38.

53 Meszaros J, Ho C. Building trust and transparency? Challenges of the opt-out system and the secondary use of health data in England. *Med Law Int* 2019; **19:** 159–81.

54 Meszaros J, Ho C, Corrales Compagnucci M. Nudging consent and the new opt-out system to the processing of health data in England. In: Corrales Compagnucci M, Forgó N, Kono T, Teramoto S, Vermeulen EPM, eds. Legal tech and the new sharing economy. Singapore: Springer Singapore, 2020: 93–113.

55 Pacurariu A, Plueschke K, McGettigan P, et al. Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation. *BMJ Open* 2018; **8:** e023090.

56 Zhang J, Sood H, Harrison OT, Horner B, Sharma N, Budhdeo S. Interoperability in NHS hospitals must be improved: the Care Quality Commission should be a key actor in this process. *J R Soc Med* 2020; **113:** 101–04.

57 Warren LR, Clarke J, Arora S, Darzi A. Improving data sharing between acute hospitals in England: an overview of health record system distribution and retrospective observational analysis of inter-hospital transitions of care. *BMJ Open* 2019; **9:** e031637.

58 Aspinall PJ. Measuring the health patterns of the 'mixed/multiple' ethnic group in Britain: data quality problems, reporting issues, and implications for policy. *Int J Soc Res Methodol* 2018; **21:** 359–71.

59 Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. *BMJ* 2003; **326:** 1070.

60 de Lusignan S, Sadek N, Mulnier H, Tahir A, Russell-Jones D, Khunti K. Miscoding, misclassification and misdiagnosis of diabetes in primary care. *Diabet Med* 2012; **29:** 181–89.

61 Boyd A, Cornish R, Johnson L, et al. Understanding hospital episode statistics (HES). London: CLOSER, 2018.

62 Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. *J Med Internet Res* 2018; **20:** e185.

63 Thygesen JH, Tomlinson C, Hollings S, et al. COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records. *Lancet Digit Health* 2022; **4:** e542–57.

64 Smart A, Harrison E. The under-representation of minority ethnic groups in UK medical research. *Ethn Health* 2017; **22:** 65–82.

65 Keyes KM, Westreich D. UK Biobank, big data, and the consequences of non-representativeness. *Lancet* 2019; **393:** 1297.

66 Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit Health* 2021; **3:** e260–65.

67 Leslie D, Mazumder A, Peppin A, Wolters MK, Hagerty A. Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? *BMJ* 2021; **372:** n304.

68 Dahlen A, Charu V. Analysis of sampling bias in large health care claims databases. *JAMA Netw Open* 2023; **6:** e2249804.

69 Ledford H. Millions of black people affected by racial bias in health-care algorithms. *Nature* 2019; **574:** 608–09.

70 NHS. NHS England—transformation directorate. National COVID-19 Chest Imaging Database. https://transform.england.nhs.uk/covid-19-response/data-and-covid-19/national-covid-19-chest-imaging-database-nccid/ (accessed Sept 5, 2023).

71 Wise J. Price hike makes access to patient data unaffordable, say researchers. *BMJ* 2019; **366:** l5305.

72 Clinical Practice Research Datalink. CPRD pricing. May 25, 2023. https://cprd.com/pricing (accessed June 7, 2023).

73 Bradley SH, Hemphill S, Markham S, Sivakumar S. Healthcare systems must get fair value for their data. *BMJ* 2022; **377:** e070876.

74 Bahmani A, Alavi A, Buergel T, et al. A scalable, secure, and interoperable platform for deep data-driven health management. *Nat Commun* 2021; **12:** 5757.

75 Zhang J, Budhdeo S, William W, et al. Moving towards vertically integrated artificial intelligence development. *NPJ Digit Med* 2022; **5:** 143.

76 Zhang J, Symons J, Agapow P, et al. Best practices in the real-world data life cycle. *PLOS Digit Health* 2022; **1:** e0000003.

77 Higgins D, Madai VI. From bit to bedside: a practical framework for artificial intelligence product development in healthcare. *Adv Intell Syst* 2020; **2:** 2000052.

78 Sheikh A, Anderson M, Albala S, et al. Health information technology and digital innovation for national learning health and care systems. *Lancet Digit Health* 2021; **3:** e383–96.

79 Bloomfield C. Sub-national secure data environment investment. Dec 9, 2022. https://www.england.nhs.uk/blog/investing-in-the-future-of-health-research-secure-accessible-and-life-saving/ (accessed June 7, 2023).