Palantir

A National Technical Framework to Underpin the UK Life Sciences Vision

Palantir Technologies
→ palantir.com/uk

December 2021

Palantir

Contents

Palantir

Executive Summary

The COVID-19 pandemic has highlighted the importance of clinical data to develop treatments, vaccines and other interventions quickly and safely. The scale and speed of the UK's response has been praised in two recent Government publications: DHSC's draft Data Strategy[1] and the Life Sciences Vision.[2] These papers underline the Government's commitment to increasing the UK's standing as a global leader in life sciences and the importance of data-driven clinical research for public health.

The UK is well placed to become the standard bearer for data-driven health research and innovation. It can lead the way in the use of data to enhance and expedite: 1. The research and development of new healthcare treatments, technologies and policies and; 2. Clinical and regulatory processes for the appraisal, approval and monitoring of such innovations. This would enable research findings to be translated into improvements in patient care more quickly and safely.

The UK can leverage key advantages: world-class research institutions; national agencies that can create a coherent research landscape (such as Genomics England, HRD-UK, NIHR, MHRA and NICE); a thriving life sciences sector; and the position of the NHS as a single payer health system with the potential to capture an integrated picture of care across the lifecycle for 67 million people.

Despite concerted efforts, significant barriers remain to realising the potential of data for health research in the UK. Data is fragmented, there are major bureaucratic constraints and the public lack confidence in data sharing initiatives.

Trusted Research Environments (TREs) alone are not sufficient. The linked Trusted and Connected Data and Analytics Research Environments (DARE UK) programme envisages that TREs (secure online environments) will be harmonised and joined-up to support and scale research.[3] However key challenges remain such as: TREs legal right to hold data; the unwillingness of data controllers to release their data for future unspecified purposes; how data quality would be improved; how the technology would actually work to integrate multiple siloed TREs into a national dataset; and public concerns related to data protection and patient privacy and confidentiality.

We propose a National Technical Framework for UK Life Sciences to bring greater coherence, clarity and control to the way health data is managed and shared across the UK. This would enable the creation of a world-leading index of data for the entire country that gives care providers and citizens full control over their data, while enabling approved TREs, researchers and regulators to access clean, harmonised data for clearly defined research and post-market surveillance purposes. Technology is now available to support such a framework in a way that wasn't possible a few years ago.

This Framework would differentiate the UK in the global healthcare and life sciences market. Most importantly, it would allow care providers, researchers, and citizens to exchange and leverage health data in the most transparent and trusted way, providing a growing data asset that would drive improvements to public health for decades to come.

Neither of the two main approaches taken to data sharing thus far have balanced the need to enhance research and innovation with the need to protect data security and patient privacy. Federated research networks cannot easily be scaled, limiting the scope of research, while centralised data lakes, such as care.data and the General Practice Data for Planning and Research (GPDPR), have caused great public concern related to data protection and patient confidentiality in recent years.

The Framework proposes a third way: a logically federated, locally controlled cloud-based data infrastructure. This would enable local data owners (care providers) to maintain control over their data, maximise the potential for research across the UK, and enable greater public transparency and engagement, including enabling individuals to opt in or out of sharing their data for specific purposes.

How the Framework would operate:

1. Patient data is aggregated locally in a virtual compartment with a standardised data structure or "ontology" that is controlled by the institution that owns the data, e.g. an NHS hospital trust or Primary Care Network (PCN), where the data is gathered.

2. The data controller in the hospital or PCN determines what goes into the institution's virtual compartment from its many clinical systems, oversees linking, cleaning and deidentification of the data, and controls what leaves the platform for any purpose, e.g. research.

3. The cleaned and linked data held in this virtual compartment is available for local use by the hospital or PCN for research, quality improvement or operational management as desired.

4. Requests for legitimate and approved research studies are submitted by researchers to the local data controller, who can take this through local approval processes. If they agree, the request is approved in the data owner's virtual data compartment.

5. The data controller selects the format in which the data is shared, i.e. whether it is shared in a pseudonymised, fully deidentified, aggregate or, with patient consent, identifiable format. The data is then compiled into the researcher's separate virtual compartment (where data can be released to their TRE) for a pre-defined period, with data updated at a pre-approved frequency.

6. Researchers have access to a cohort builder tool, enabling them to know what data is held by each data controller and design their request without knowing the contents of the data.

7. Data supplied from different data controllers is aggregated within the researcher's own virtual compartment, creating the data layer of their TRE onto which they can lay their preferred research analytics tools. This aggregation of data from different sources is enabled by the fact that all institutions use a standardised ontology.

8. Data releases are audited and fully traceable, and once the agreed purpose for accessing the data expires, data sharing ceases. When the researcher has finished their research, the data will be locked, so that it cannot be used for unapproved studies, but can be unlocked to be analysed by future researchers if needed.

9. Data use is transparent, enabling the public to see and control how their data is being used. Citizens will be able to opt in or out of their data being shared for different categories of research.

What the Framework would deliver for stakeholders:

This paper is the result of many conversations with national and international stakeholders and experts, as well as Palantir's direct experience supporting leading collaborative research environments like the National Institutes of Health's National Covid Cohort Collaborative in the U.S. We welcome additional feedback to further develop the proposed framework.

Palantir

The Challenge

The COVID-19 pandemic has highlighted the importance of access to clinical and population data to develop new treatments, vaccines and other interventions quickly, effectively and safely. The UK is justifiably proud of its role in the response: it must now act to capitalise on the progress it has made.

Data-driven clinical research could bring immense benefits to the NHS and the people it serves — from enabling better prevention and early diagnosis of disease, to developing more tailored and effective treatments and speeding up the roll out of cost-effective innovations, in support of the NHS Long Term Plan[4] to deliver more preventative, personalised and integrated care and address health inequalities across the UK.

However, there are tensions between the priorities of researchers and life science institutions and those of care providers, patients and regulators. Notably, efforts to realise the potential of health data to drive improvements in care must be balanced with the need to engage the public and protect patient privacy.

Researchers and life sciences institutions want:

→ Ease of access to data, requiring simpler governance and oversight of NHS data with smoother and faster access to Real World Data (RWD).[2]

→ Discoverability of datasets, with a maintained list of trusted data providers that tracks individuals' data across sources.[5]

→ Assured data quality, making all types of data research-ready and reducing the time that it takes to conduct research and produce results.[2, 5]

→ More effective and efficient clinical trials, through larger scale data, faster cohorting, and repeatable analyses.

The public, providers, and regulators want:

→ Data access that is proportionate and aligned to defined purposes, in accordance with project requirements agreed with patient and public representatives and in line with GDPR.[6]

→ Data access that is transparent and trustworthy, with sharing only allowed for legitimate purposes and in compliance with patient consent guidelines.[2]

→ To understand the benefits of data sharing for public health outcomes, patient care and tax payer value for money. The social and financial returns of data sharing initiatives must be made clear. This incudes NHS or Government investments in resouces to enhance data management and sharing capabilities, as well as equitable partnerships that involve the sharing of data for fair use.[7]

→ The tools to deploy research discoveries into practice and monitor outcomes, so that new medicines, treatments and care approaches can quickly and equitably be made available to patients across the NHS. Outcomes must also be tracked to check whether impact in practice matches reseach findings, and identify any issues or opportunities for improvement.

Attempts to resolve these tensions have not yet been successful at scale. Neither of the two main technical approaches for using health data for research – federated research networks and centralised data lakes – fulfils all of the requirements outlined above. Federated research networks have limitations for scaling, thereby limiting the scope of research, and centralised large-scale data lakes have caused great public concern related to data protection and patient confidentiality in recent years. (These challenges are set out in more detail in section 2.1 below.)

Public trust in data sharing has also been undermined by recent initiatives where there has been a lack of clarity over what data is being shared, how it will be used and how patient confidentiality and privacy will be protected. This includes efforts to aggregate data into centralised "data lakes", exemplified by national programmes such as care.data and the General Practice Data for Planning and Research (GPDPR), as well as national and local deals to share cohorts of data with individual companies.

Trusted Research Environments (TREs), also known as a 'Data Safe Havens', have been developed as a technical solution to enable data controllers to share data more securely, whether through federated research networks or centralised data lakes. TREs are defined by NHSX as "controlled digital environments used to store or analyse sensitive data securely" and to "support their standardisation".[8]

A large number of TREs have now been created at national and at regional levels, for example through the Digital Innovation Hubs that made data from some of the major health providers accessible.[9] Unfortunately, this means we now have a fragmented data landscape where data cannot be optimised for research and innovation in the NHS or life sciences sector. Researchers cannot easily understand what data might be available that is relevant to their purposes. Data also cannot easily be shared between TREs, which limits the scope for analysis of related datasets.

We believe the UK needs a national data sharing infrastructure that enables the NHS and researchers to work in partnership, with clear mechanisms to protect patient privacy and data security. This is the key to enabling health data to be used at scale whilst maintaining the trust of the public and care providers.

Such is the fragmentation that there are now calls for interoperable TREs[2] with common data standards for data sharing and developer-friendly efficient aggregation of datasets.[5] A similar approach has been attempted with the establishment of Local Health and Care Record Exemplars (LHCREs), which aim to develop shared health and care records for people in a particular region for better coordination of care. However, LHCREs are not operating to common data standards or routinely sharing data when patients move to another region, or receive care outside of the LHCRE where they live, which demonstrates the difficulty of achieving integration in practice.

Integrating TREs also requires more than the ability to share data easily. There must be clear mechanisms for maintaining patient privacy and data security when data is shared. The UK Health Data Research Alliance has set out a list of features that TREs should include.[6]

Current Approaches

Currently, there are two main approaches to using health data to facilitate clinical research, neither of which adequately balances the priorities outlined above.

① Federated research networks, whereby data is stored on physically separate infrastructure and the same analyses are run on each institution's data independently. The federated approach makes it easier for institutions to maintain control over their data and protect patient privacy, but it has a number of important limitations:

- → It uses static datasets, rather than real-time data feeds, which means that data quickly gets out of date.
- → It is harder to detect and resolve data quality issues and biases across the multiple datasets.
- → It limits the scope for running statistical models and machine learning algorithms, because the data is split into smaller data sets, thereby limiting the size of training data available.
- → It requires time-intensive data preparation, because data structures must be identical at each institution so that results can be compared. This may require local data experts to conduct complex, time-intensive data transformations – work which has to be repeated if the data is used for a different study or purpose.

② Centralised data lakes (national and regional), whereby data from different sources is integrated into a single asset that researchers can access. This addresses many of the technical and operational issues associated with federation, but makes it more difficult to control data access, protect patient privacy and ensure data security.

The concept of creating a giant data lake that researchers can access as and when they need to also conflicts with the core information governance requirement that data should not be aggregated without a specific lawful purpose. Recent attempts to create a national health data asset have tended towards this "giant data lake" approach. Unsurprisingly, this has caused concern among privacy campaigners and the wider public.

Regional data lakes may be less contentious than national ones, and allow local communities to exercise stronger oversight, but they face similar data protection and privacy challenges. This is because they still require a high degree of harmonisation and integration of data from different sources to fulfil varied research needs within a region, as well as planning and other purposes.

Regional data lakes would also need to operate around different geographical and organisational constructs depending on the purpose of sharing data. For example, research needs might surround an Academic Health Science Centre, operational reporting needs for benchmarking might operate around Integrated Care System (ICS) boundaries, and shared care records are increasingly operating across many ICSs, as initiated by the LHCRE programme.

In addition, existing regional TREs (many of which currently persist data as a data lake), do not offer equity of opportunity for research, as they tend to centre around large prestigious teaching hospitals. Smaller care providers, such as District General Hospitals, often lack the resources to participate in research, despite being keen to do so. This means that smaller institutions miss out on the potential benefits of research partnerships, and researchers cannot access relevant data that such organisations may hold.

The UK's various TREs have offered some centralised and linked data to researchers, governed by the ONS Five Safes framework.[11] This framework sets a necessary baseline for data protection, but it is not sufficient for robust scaling of the use of data for different purposes. It does not in itself indicate how to create the right research infrastructure in the UK.

The growing number of specialised TREs do not have a common specification or mechanism to enable shared purpose-based access to data, restricting the kind of large-scale analyses that are often required for effective research. Collectively, they are currently unable to provide a single national index of all data available for research, they fragment the ability to audit the use of NHS data for research and provide transparency to the public, and each will separately tackle data quality rather than improving it once for all. HDR UK has been working with others on metadata model harmonisation across TREs and with ICSs, with the intention of making data more easily discoverable.

The Government's Life Sciences Vision proposes accrediting a handful of TREs as the default route for researchers to access large-scale NHS data,[2] a process which could be facilitated by the recommendations made in this paper, and harmonisation work for TREs is progressing with the Health Data Alliance, HDR UK TRE Working Group, the DARE UK programme and new NHSX TRE Reference Groups.

The Department of Health and Social Care (DHSC)'s draft Data Strategy[1] sets several objectives relevant to TREs, including:

1. Giving health and care professionals confidence in sharing data. Many staff are concerned about breaching confidentiality when using data and find information governance rules difficult to apply. In our experience, this particularly affects front-line clinical staff, who are not receiving the holistic data they need on their patients and end up repeatedly collecting the same information as a result. Innovations in security controls would improve clinicians' trust in the system and their confidence that data will only be shared for approved purposes.

2. Increasing the scale of research-ready data. The Data Strategy aspires to the richer insights that can be gained from large-scale data. However, the analysis of health data at scale is often impeded by inefficient federated research methods and the proliferation of TREs.

3. Improving the quality and accessibility of research-ready data. The Data Strategy acknowledges varying data quality in organisations, and defines the need to validate data at the point of entry to improve quality and make all data discoverable.

4. Giving people the ability to access their health data and see how it is being used. The Data Strategy commits to giving citizens the ability to see what research their data has informed and who has had access to their data.

5. Closing the gap between the lab and the delivery of care. Research environments should create a loop in which research data and results are made available to clinicians to benefit patients at the point of care, and RWD is routinely collected from provider settings and shared with researchers.

The federated research network and centralised data lake approaches described above are inadequate for meeting these objectives.

The federated model only addresses points 1 and 3: giving health and care professionals confidence in sharing data and improving the quality and access of research-ready data. In addition, achieving either outcome is dependent on the way the federation is managed.

Federation keeps the data local, so there is the potential for local ownership, curation and supervision by the clinicians who are gathering the data, but only if the federated dataset remains within the NHS for research. If the federated data is transferred to a research institution, it forms a collection of small data lakes that are out of NHS control and are dependent on local research practice to ensure they are not used beyond the approved purposes.

The centralised national or regional data lake approach only addresses the challenge raised in point 2: increasing the scale of research-ready data. It also disconnects the original data controllers from their data and requires the aggregation of large quantities of data with legal purposes yet to be defined.

A New Approach

This paper proposes a National Technical Framework that scales from local to national and addresses the objectives laid out in the DHSC's draft Data Strategy (outlined in the previous section), without compromise.

Technology is now available to support this in a way that wasn't possible a few years ago. There is an opportunity to develop an alternative model that maintains local control while maximising the potential for research across the UK. This paper sets out a new National Technical Framework to facilitate this model: a logically federated, locally controlled, cloud-based data infrastructure.

This Framework would maintain local control for data controllers and support the cleaning, aggregation and onward usege of data for clearly defined and lawful use cases. Under the right controls, high-quality data could be shared for wider purposes internally or with partners. This would enhance research, innovation and the running of the health and care system at all levels.

This approach recommends the separation of the key functions that each TRE might otherwise perform, such as data storage or anonymisation, and layers in additional controls that support information governance requirements, whilst enabling public engagement and trust in a national-scale architecture.

The following section lays out the requirements for the Framework, with a case study that outlines how such an approach could operate in practice. The paper then summarises the benefits of the Framework for research and the future opportunities for enabling the Life Sciences Vision.[2]

A National Technical Framework

In order to achieve the Government's strategic priorities for life sciences and resolve the tension between the use of data for research and data privacy, we recommend a National Technical Framework with the following features:

01 **Curation of high-quality data at source.** →

Local ownership and curation of patient data in identifiable format for each legal entity (Hospital Trust, Community Trust, Mental Health Trust, Ambulance Trust, Primary Care Network for General Practice, etc.) should take place in a logically separated virtual compartment. This would be a cloud-based data compartment that is owned and fully controlled by the legal entity and can be used for its own purposes.

As part of the overall Framework, the virtual compartment would need to comply with the information governance and access control features described below. The legal entity would be responsible for improving data quality, with data quality tools provisioned across sites, so that quality issues discovered by wider users could be fixed once for the benefit of all users.

Data in the compartment could be transformed once to research-ready formats such as the Observational Medical Outcomes Partnership (OMOP) common data model. Data controllers would be incentivised to create comprehensive, high-quality datasets, because they could use the curated datasets for their own purposes, such as quality improvement, audit, or operational management.

02 **De-identification at source.** →

Data controllers must have the ability to anonymise and pseudonymise their data in their own virtual compartments. It is desirable that such work is only carried out in accordance with the Framework to ensure consistent standards are applied, with the ability to audit data as well as prevent the need to release identifiable data to TREs. Data controllers would also be able to create synthetic data to enable innovators and researchers to build applications and models without the need to start with patient data.

03 **Purpose-based limitation on data sharing.** →

Local data controllers must maintain control over the decision to share defined data for any lawful and approved purpose, including their own usage of data. This can be accomplished through purpose-based access controls, whereby external researchers, regulators and other parties can only gain access to defined data by issuing a purpose-based access request. These requests should be digitally approved, with permissions set to persist as the data is used, and all actions logged for auditing.

04	Flexible and reusable ontology layer.	→	"One man's data curation is another man's data destruction"[12] – meaning that in the process of harmonising data for one purpose we can destroy its ability to be used for other purposes. For example, harmonising data in shared care records for population health management would enable segmentation, risk stratification and identification of gaps in care, but would not contain the granularity required to run an observation study for a complex medical device. Therefore, the Framework should have a flexible ontology layer with data modelling tools that enable research organisations and end users to build ontologies within the Framework based around their study criteria and share these models for others to use. Researchers should be able to access these data models to define their data requirements without accessing any patient data, prior to submitting a request for purpose-based access controls. This would enable them to define the data they need to request. The ontology layer should adopt open standards such as the metadata model standard being developed by HDR UK.
05	Near real-time availability of data.	→	This will be required to support a wide range of current and emerging MedTech use cases, for example digital diagnostics and digital therapeutics. AI/ML tools require incremental data to learn and recalibrate. They need access to data at each local level to deal with context-sensitivity, as well as an automated way to track performance of the algorithms and models against RWD, so they can address calibration drift or failure. Near real-time access to data would also offer opportunities to enhance regulatory oversight of medications and treatments, with data driven post market surveillance and more efficient monitoring and reporting of performance, including adverse events and reactions.
06	Minimised and effectively controlled data duplication.	→	When data controllers need to copy their data, they should have the ability to do so in a way that tracks and version-controls any duplicates and gives visibility on how they are being transformed and used. This offers a compromise between the architectural principle of zero duplication of data for planning and research uses, and the reality that health data is often already duplicated across multiple source systems. It also means that duplicates would only be made when necessary and for a justifiable purpose, and that the copies would be auditable and quality controlled by local staff who are best placed to interpret the data.
07	A nationally indexed data catalogue for discovery and cohort building.	→	The Framework should provide a data catalogue that indexes all data available through the logically federated compartments, with accessible metadata and quality metrics, and should utilise HDR UK's work on metadata harmonisation. This would help create a fairer environment for data partnerships with the NHS, since high-quality data from smaller Trusts would be equally as discoverable and usable as data from the larger urban trusts that have been the traditional targets for high profile partnerships.[7] For study creation, it would offer tools to enable researchers to easily build cohorts without being able to see individual patient data, and explore data linkage across anonymised datasets prior to data release.

08	Time-limited data release to approved research or analytical services.	→	Central data marts should be created in the Framework from the local data sources and released for analysis to any approved tools chosen by the researcher. This should exclude patient data where use has been prohibited by the patient through their consent controls (such as the National Data Opt-Out). Nationally curated standard datasets and researchers' own data could be linked to research data marts prior to data release. The data mart should be regularly refreshed according to the research protocol and made available for a fixed period of time only. At the end of the study, the data mart should be locked in the Framework so that it cannot be used beyond the approvals given, but it could be unlocked as agreed for reproducing or continuing published studies. If there is no need for the chosen research or analytical tools to retain this data then it should be removed, minimising data duplication.
09	Controlled governance and public transparency.	→	By default, the Framework should hold a record of what data was used and from which sources, as well as which research and analytics services it was shared with for which approved purposes, ideally with confirmation from those end services when the data has been removed from their environments. This would provide transparency and the ability to audit how data is being used across the UK. It would support the creation of a central register for data partnerships, with visibility on their data usage and what associated research benefits they are bringing within specific collaborations.
10	Dynamic consent and patient control.	→	Since data would be tracked with full provenance and auditability, the Framework would also offer complete transparency for patients on how their data is being used. With a simple Application Programming Interface (API) service, citizens could access this information from their NHS App, and set consent preferences for future use of their data that is administered in the Framework once for all sources. This should go beyond National Data Opt-Out preferences, and provide a capability that can integrate with the NHS App to give citizens more granular control, for example: → Opting in or out for identifiable or pseudonymised data by category for different types of study, such as selecting 'Yes' to sharing data on acute and long-term physical health conditions but 'No' to sharing mental health and sexual health data. → Setting permissions for when data can be used without further consultation with the citizen, when they should be asked if they want to opt in or out, or when data should never be used. → Accessing a register showing citizens where their data has been or is being used and where their permissions have stopped it being used, so that they can change their permissions if they wish. → Notification features, so that citizens know every time their data is used and what it is being used for.

A suitable governance structure will be required for the Framework, including an oversight board to monitor its function, development, and operation. The board would benefit from being independent, especially if it were to own pseudonymisation keys or processes, like re-identification approvals, and work with the public to agree and oversee controls.

Palantir

Further Considerations

A range of important additional points to consider in developing the proposed Framework were raised by stakeholders consulted on this white paper. We have outlined these considerations below, some of which will require further stakeholder consultation and collaboration to address.

How much data from each data controller?	→	A key consideration for the Framework is what amount of data from each data controller organisation would be available in its virtual data compartment. This could be decided by each organisation based on the data it wants to make available, but would likely need national guidance on minimum data sources and datasets to support a large range of research. Trusts such as hospitals operate hundreds of information systems, not all relevant to research, innovation, and planning, and could at their discretion make additional data available for research, beyond any minimum requirements. The scope of data research and innovation use cases that arise could spur the need to make more data routinely available in the Framework.
Patient level data from research organisations.	→	The Framework could also connect patient level data from clinical trials and other research to facilitate further research or even to inform clinical care. To enable this, participating organisations would have their own logical virtual "sending" compartment to make data available to other organisations, and control access in the same way as care providers, through purpose-based access controls. When studied only 25 percent of large pharmaceutical companies met data sharing standards.[13] There is an opportunity to engage industry in sharing it's data in return for access to NHS data. Attention would need to be paid to ensure that any data that is shared is covered by informed consent mechanisms. Further work is needed to explore how this could work in practice.
Handling -omics and imaging data.	→	Considerations will need to be made for the technical challenge of handling -omics and high-density imaging due to their large size. Decisions will need to be made on how much of this data should be routinely available within the logical compartments of the Framework versus brought into the Framework on a needs basis, or whether it should always reside outside of the Framework and require separate controlled access. This could include consideration of what level of data is available, for example the Framework could routinely store genomic variation data in the Variant Call Format (either once nationally or separately stored in each federated compartment), but not whole genomes, which would be stored in specialised genome stores.

Responsibilities for improving data quality.	→	Data quality is the most important topic if data is to be valuable in supporting research and innovation. Poor data quality is often the elephant in the room in discussions on how to share data for such purposes. Whilst every data controller can and should be responsible for data quality, there are significant opportunities to improve data quality in this centralised Framework. It is difficult to discover site-to-site differences in data quality using federated checks alone, because the data cannot be cross-referenced between sites. In contrast, centralised data quality benchmarking could reveal unique opportunities for data quality improvement that would support improved research analytics locally and in aggregate.[14] Central tools could help establish non-conformance of data-to-data models, density or completeness and use frequency distributions to determine plausibility. Some data quality issues could be auto corrected for onward usage, but they should be flagged in the Framework back to data contributors to ensure corrections are captured by the data owner within its virtual data compartment, for the benefit of all future users. It could be the responsibility of the researchers and innovators that are using outputs to use such central data quality tools and other tools to flag individual row level data quality issues back to data controllers.
Resourcing and workload.	→	It will take effort to manage this Framework at a local level with the potential for significant distributed governance workload. Technology could facilitate this, e.g. purpose-based access requests could be delegated to functions. A busy GP will not be able to handle 30 access requests a day, but they might delegate this responsibility to administration in a PCN or ICS. Categorising purposes and creating and evolving governance around them would help to determine which types of use cases might be automatically approved with audit trail (e.g. MHRA post market surveillance of a new drug), and which require explicit approval (e.g. research studies requiring patients to be identifiable). Each data controller will need to invest in improving and maintaining data quality, dealing with automated and manual alerts on data quality issues from downstream users and research programmes. They will need to put in place processes and investment that are likely to be commensurate with the gains for doing so, for example increased success in grant applications, increased revenue from commercial sector for running trials and using data. National bodies must consider the local investment in data that will be required to realise the Life Sciences Vision and make appropriate investments.
Consistency of purpose approvals.	→	We anticipate variation across care providers in whether they approve purpose-based data access requests. In some cases, this will be of no consequence, for example a research programme looking for volunteer care providers to participate. In other cases, it could have a profound impact on the success of the proposed research, for example a study on a rare disease, where potential participants are limited. Regions are putting in place governance for research and decisions could be made at this level to ensure consistency for specific programmes. This could result in care provider decisions being delegated to such bodies transparently.

Purposes that citizens can understand.	→	For NHS research and other legitimate uses of data where explicit consent is not required, a pragmatic approach to engaging citizens will be required. For example, anonymous data could be used from a patient in a mental health study, but the patient themselves might not have a mental health condition. Communicating this purpose could be confusing or worrying for the individual concerned. It is likely that we will require a standardised approach that outlines understandable categories for patient data use against which citizens could give permissions. The following examples were provided by Professor Dipak Kalra from the European Institute for Innovation through Health Data, which is actively working to provide standardised approaches:[15] → Categorise purposes that the public understands and can make decisions around e.g. drug development or public health, and record preferences against these purpose categories. This will also make it easier for patients to manage opting in and out of data sharing. → Enable patients to see examples of research that has been conducted within categories to help them understand decision making at the category level. → Enable patients to profile their preferences against these purpose categories, e.g. a patient might choose to opt out of their data being used for AI/ML. → Enable care providers and their data intermediaries to engage the public on what purposes they will support and what purposes they will not support e.g. developing a stance on tobacco research. This would enable organisations to signal how they will and will not use data.
"Consent for consent" processes.	→	How does a researcher access someone's personal confidential health data lawfully and ethically in order to identify whether they would be suitable for a research trial and then to contact them for their consent? This is the "consent for consent" conundrum. The usual practice is to use anonymous and aggregate data to identify how many patients might be eligible within a care provider, so that the care provider can legitimately reach out to the patient to ask if they consent to being approached by the researcher. If an individual consents to this, the researcher then contacts them to consent them into the research study. This "consent to be contacted" process is also used in some research studies based on analysis of existing data. However, such studies can and do use the Control of Patient Information (COPI) regulations to set aside the common law duty of confidentiality via section 251. Both processes are unsustainable and unsuitable for an enhanced scale life sciences industry.

One solution is to change the law to allow permissive processes with or without the counterbalance of allowing individuals to opt-out. This could be undertaken nationally or regionally, for example the cross sectoral draft European Data Governance Act allows for data intermediaries to form broad consent arrangements with patients/the public to have access to data and make it available to others without requiring explicit consent (it is unclear how the Act would apply to the UK post-Brexit).

The NHS could be one such intermediary and form a broad consent arrangement with the public as part of its future constitution. An alternative could be community owned data intermediaries that return profits to the community they serve. Any such arrangement would need all the controls articulated in this Framework in place to govern its work.

In the absence of a change in the law, what could be done to break through the current consent for consent constraint? The NHS already supports research and the use of private sector technology for direct and indirect care through Trusts and GP practices. What has failed to materialise is one or more suppliers providing a compelling offer to NHS Trusts to enable them to engage with the life sciences agenda other than within the current mechanisms as laid out above.

There is also a disincentive for the NHS to take a strategic decision in this direction as the National Data Opt Out (NDOO) rate may increase, which is seen as a negative outcome even though we know that about 20 percent of people have concerns about the use of their data for planning and research uses.[16] A compelling case could be made for national NHS research, by which we mean:

- → The research is governed by Health Regulatory Authority. It is ethically approved by an NHS LREC or MREC; the protocol is included on the NHS Research Register; it is costed using HRA methodology; and it conforms to the HRA governance standards.

- → Patient participation is by consent only. Patients receive a two-step contact:
  - ↳ Step 1: Statement of eligibility for research and a lay summary of research. This results in a simple positive action for patient to accept or decline contact by researcher. No response means no consent to be contacted.
  - ↳ Step 2: Researcher makes contact with patient via email or text.

- → Patients can dissent at any point in the process and can opt out of the whole process via the NDOO.

- → There is an explicit financial model open to public scrutiny in which the patient could gain from their data financially or choose for that money to be directed to NHS funding.

This would not cover all types of research but would be a sound place to start, before moving to purely commercial, policy or other non-NHS research. It also does not obviously conflict with existing or proposed legal changes.

Stakeholder Experience

The National Technical Framework would enable operational value for both data contributors and consumers, as well as the public at large, in an environment built around privacy and mutual trust.

Figure 1: A National Technical Framework can benefit all stakeholders, giving different groups the data they need transparently and with flexibility in usage and choice of research tools.

Framework Architecture

The Framework architecture would consist of three layers, as shown in the diagram below:

❸ Research, analytics and MedTech services

❷ Data Discovery, Provenance, and Governance Services for purpose-based access

❶ Data Integration & Curation Services

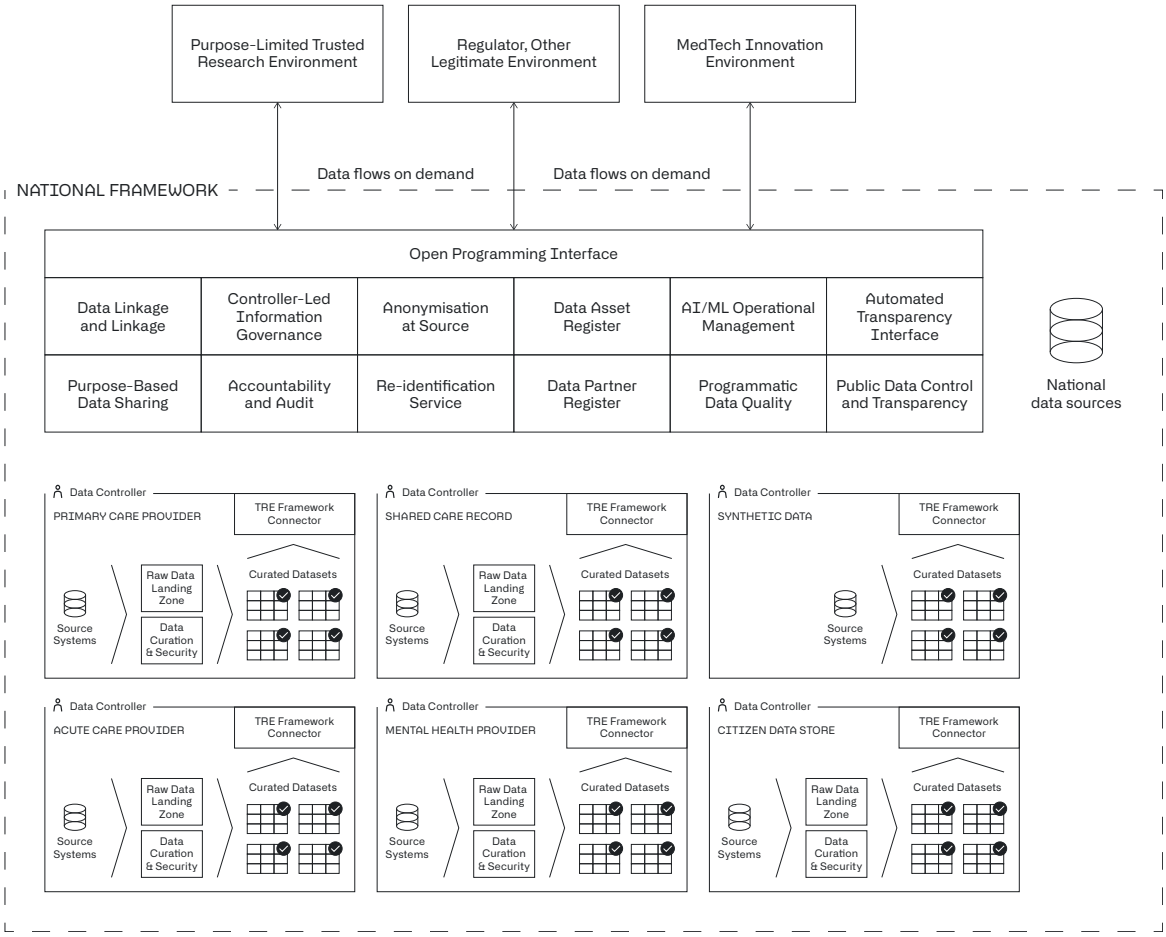


Figure 2: A National Technical Framework can be achieved through a three-layer architecture, delivered using one or more technology solutions that can provide the compartments and tools needed in each layer.

Layer 1: Data Integration and Curation Services. This provides logically federated virtual compartments for each organisation that contributes data, allowing them to maintain total control of their data. In this environment, contributors can clean, pseudonymise, de-identify and transform their raw data, apply any necessary privacy and access controls, and perform quality checks to ensure data is ready for broader use, whether for research or other purposes. Data can be made available incrementally and in near-real time to support MedTech innovations that require this frequency of data.

Data controllers can grant purpose-based access to the transformed datasets as they receive requests from researchers or any other organisation, and audit downstream data usage. This approach also creates a logical data layer that is separate from applications, as recommended in the DHSC Data Strategy.[1]

Synthetic data could also be held by each data controller or could be provided as a separate compartment, for example for linked synthetic data, which would require its own ownership and governance. It would be beneficial for innovation to have competing synthetic data compartments that could end up being optimised for different research and development purposes.

Layer 2: Data Discovery, Provenance and Governance Services. This offers horizontal capabilities that can be applied to all data contributed for research, to make current, cleaned data accessible to researchers and innovators. These capabilities include:

→ Live data connections to local data compartments, so that researchers and innovators can access data in near real time.

→ Data lineage tools, giving researchers visibility on how data has been connected and transformed so that they can explain this in their study write up.

→ Data linkage services to join data across disparate datasets.

→ A data asset catalogue, which can be public-facing and used during the process of trial design.

→ Data modelling tools, enabling research organisations and end users to build ontologies within the Framework based around their study criteria, and share these models for others to use.

→ Operational environment to operate AI/ML and other MedTech tools that require algorithms and models to be continuously evaluated against real world data to enable deterioration in performance to be identified and flagged to authors to address.

→ Programmatic data quality controls, enabling automated data quality checks for researchers and innovators to manage data quality.

→ Data protection features for local controllers, allowing them to minimise data access and set access controls according to research requests.

→ Data release and syndication following approvals from participating data controllers, to take data into the chosen TRE, research or analytical tools and use it for the requested purpose.

→ A re-identification service under the control of the Framework's oversight board, allowing critical research findings to be translated into patient care.

→ Public data control and transparency tools to allow citizens to set preferences, control their data, see how their data is being used and receive notifications.

Layer 3, Research and Analytics Services, comprises market TREs, research, and analytical tools that can receive and onwards-control the data for the duration of a study. Each organisation operating these services should have its own controls to time limit and restrict the use of data for approved purposes. Layer 3 solutions already differ in the market. Some are provided through industry partnerships and others by Government agencies. Many are specialised, for example in the field of genomics. The Framework will enable an open market for these services, which could be provided by any organisation or approved technology once the two foundational layers are in place.

Additionally, Layer 3 will enable regulators, arm's length bodies and other services with lawful data processing rights to take advantage of the Framework. Many organisations such as NICE and the MHRA will require access to RWD to evolve their services to support life science innovations, while Trusts and other NHS organisations could leverage the Framework as a clearing house to automate central returns of data. In such instances, the Layer 3 end points could be general analytics platforms rather than TREs.

The Framework could also support more advanced and evolving MedTech needs, such as organisations providing AI/ML algorithms and models, digital diagnostics or digital therapies. Such organisations could integrate with Layer 2 services so that innovations could be operationally managed against RWD to monitor performance. This would allow innovators to keep on top of complex safety issues such as context-sensitivity and calibration drift.

When research purposes are closed, the data could persist in the national Framework to support continuous improvements through the contribution of real-world evidence. This would enable the move to living guidance as outlined in the recent NICE five-year strategy,[17] and provide the additional data required for contingent approvals for the new drug and device innovations that are rapidly entering the market.

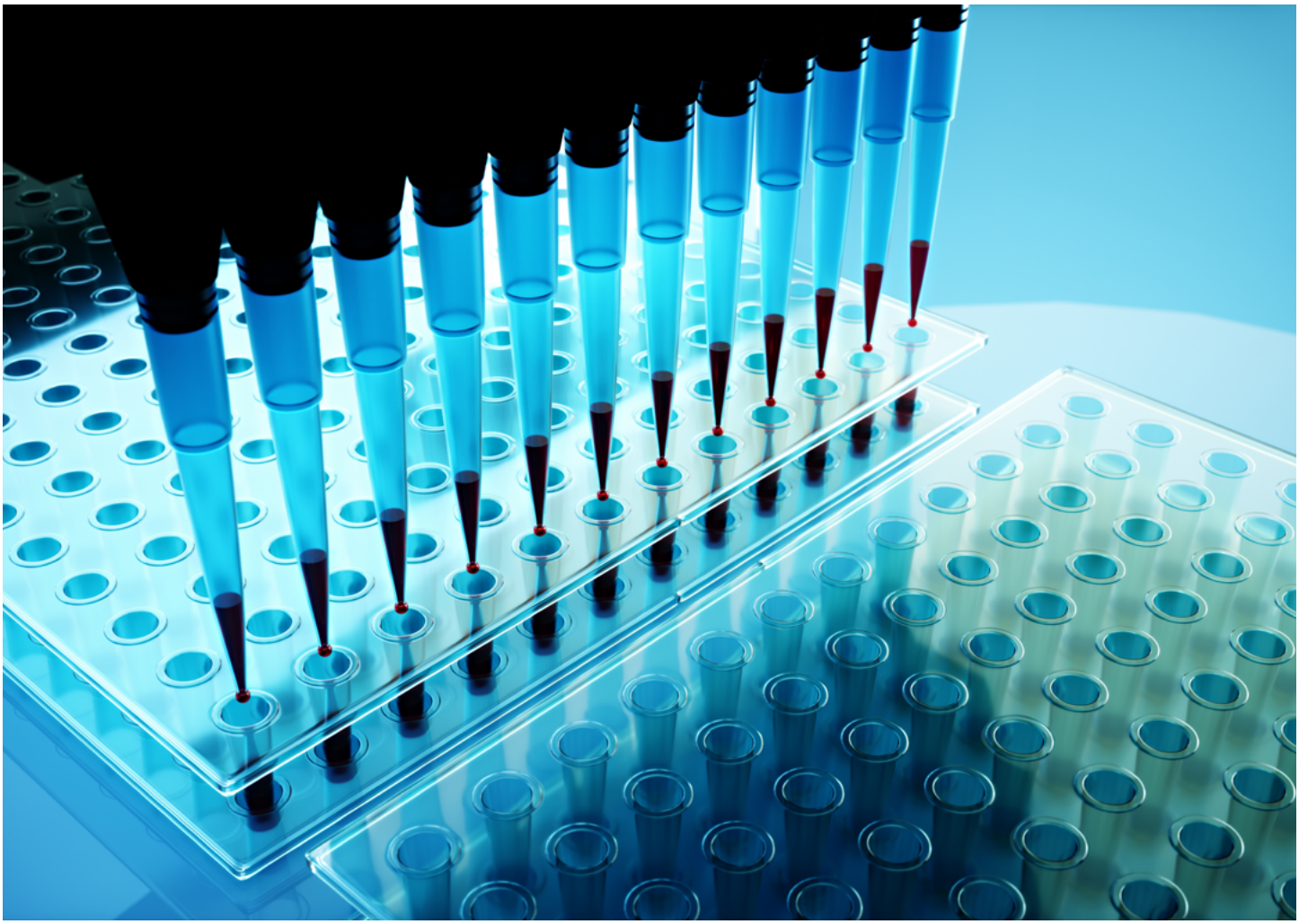
Palantir

Benefits of a National Technical Framework

The National Technical Framework has a number of advantages:

Faster discovery and access to data for researchers.	→	The Framework would enable the creation of an index of all research data available across providers which, if linked to the current Innovation Gateway to incorporate existing TRE data, would provide a single source for researchers to search and assess the feasibility of their studies. Complementary search and cohort building tools would accelerate the grouping of anonymised subjects at a breadth and depth appropriate to their studies.
Drastic reduction in effort for data quality and harmonisation.	→	Data extraction from source would only be configured once and be under the complete control of the local data controller. This improves on the current approach, whereby every research or analysis environment requires its own data extraction from the same data sources. It would enable data quality to be rapidly improved once for each data source, speeding up research and enabling quality assurance.
Minimal data duplication.	→	The Framework would minimise multiple copies of data existing in multiple locations and the inherent risk that is associated with the current federated landscape. The data would only reside in the Framework solutions or on a temporary basis in Layer 3 services, unless longer term persistence were warranted.
Simplified purpose-based access to data.	→	This would improve the ability of data controllers and the NHS as a whole to ensure that they are complying with GDPR and other legislation and regulation.
Levelling up research and innovation across the country.	→	The Framework would allow monitoring and reporting of research and innovation activity across the NHS. This would help target improvement in places where the focus on research and innovation could be increased, and support increased uptake of clinical trials in underserved communities and population segments, as called for in the Life Sciences Vision.[1] Enabling all data to be indexed would help smaller and more rural healthcare providers participate in research when they may not otherwise be able to. It would also support an equity of research so that all cohorts are represented, helping to address health inequalities.

Auditing and governing data partnerships.	→	National and public transparency of data usage by data partners would enable a national register of data partners and the auditing and evaluation of these data partnerships for fair value exchange with the NHS.
Standardisation and interoperability of research and analytics platforms.	→	The separation of the data through Layers 1 and 2 would resolve current problems of data quality and security, lowering the bar for interoperability for research and analytics tools in Layer 3. The Framework would also enable a competitive marketplace for the growing number of research and analyical tool providers, allowing them to focus on where they deliver value and compete on providing specialised analytical and modelling capabilities to researchers. This interoperability and choice would help to avoid lock-in for data controllers and research institutions.
Real-time access to data for planning and research purposes.	→	The Framework could operate in near real time, allowing a range of use cases for the planning and research use of data. This will be increasingly required by MedTech companies and regulators.



Worked Examples

This section provides some hypothetical examples of how the Framework could operate in practice to help bring it to life.

Patient	→	I was recently approached about joining a clinical research study with a pharmaceutical company. I have been approached for clinical trials in the past given my medical condition is of interest to researchers. What was different this time was that I received an email about it rather a clinician calling me, and I was given the option of whether I would prefer to receive the payments for my participation or whether this could go back to the NHS.
	→	I am able to see via my NHS App how my data is being used by the NHS to support its research and operations. I took a look at the App and saw that I could set permissions around research categories that I would prefer the NHS didn't use my data for, like use in AI, or data categories I don't want shared, like sexual health, and could opt out entirely of any data use if I so wished.
Clinical researcher	→	I am a rare disease researcher and access to national data is essential. It has been hard for me to obtain data from so many different places in the past, which limited the research potential.
	→	I already had access to a TRE with all the analytics tools I needed, including the ability to analyse genomics data, but I needed easy access to data from across the country.
	→	I was able to write my research project data needs using a tool that is connected through the ontology to the sources of data and cohorts available and I could do this without access to the underlying patient data.
	→	Through the national tool I could see an index of data for the country, and I could use this to target the datasets I needed from specialist centres that look after the types of patient I needed in my study.
	→	After I had followed the usual procedures for approval for my research requested within the tool, I was also asked to complete a form for purpose-based access to the data where I had to describe the study, including a lay person's description, and assign it to a category of clinical research.
	→	Over the coming days I was able to track the approvals coming in from the specialist care providers across the whole NHS that I had targeted to participate in this research.

	→	From here I was able to request the release of the data into my TRE for the duration of my study.
	→	I was able to save and share my cohort definition, which is useful as I am likely to come back to this cohort for future research and I know some of my colleagues around the country and world would also benefit from this.
	→	Once I received the data within my TRE I could begin exploring and analysing it. I found in the course of my work that there were some significant data quality issues with some of the sites. I was able to go into the national tools provided and flag these issues back to data quality teams at the originating organisations, so they could correct them.
Hospital provider	→	I run a research team in a small hospital and am glad that we now have tools that help us participate in wider research programmes, something we were struggling to do before.
	→	We received some funding and support from a national team to get set up and we put in place a small local data quality team responsible for improving our data quality, for the benefit of ourselves others.
	→	We have overseen the work to convert our data to the OMOP format for research, and have also established a Fast Healthcare Interoperability Resources (FHIR) version for use with App innovators that want to work with us. The process of transforming the data into both formats was accelerated by using templates created across the Framework. We also set up an anonymised and a pseudonymised version of the data to be used for reporting and research.
	→	We have gone from a position where we were hunting for research opportunities to receiving a number of purpose-based data access requests a week. Some of these requests have been from our own Trust, where I would not normally have had visibility in the past, but most are clinical research requests from elsewhere in the NHS, and one has been from a pharmaceutical company that is targeting a new clinical trial in our area.
	→	We have a governance committee that meets weekly to discuss and approve the requests that come through, and we have established rules for accepting some types of requests automatically without the need to discuss them further.
	→	I can now see all the ways in which our data is being used and am able to track this. I can report this back to our executive management team, which values this transparency, and we review data usage in our audit and risk meetings. We hope to use this information to understand what innovations we have supported and to help attract more research and innovation to our area.

	→	We have now started to receive notifications on data quality concerns through the national tool and this helps guide our small data quality team on where they need to focus their efforts. We work with our clinical and departmental teams on the data quality issues that come up and have started having a regular meeting with our internal electronic health record team, which sees opportunities to improve data capture in the user interface.
MHRA post market surveillance	→	I work in a post-market surveillance team, and we have recently embarked on a data project for a new drug that has been bought nationally by the NHS in a population health deal.
	→	Historically we have waited on research teams to conduct observational studies, receiving data through yellow cards or drug companies notifying us of issues. In this case, we took advantage of this new national tool to use a data driven approach to post-market surveillance in the market.
	→	The drug is a new innovation with huge potential and therefore had a rapid approvals process. We want, however, to help mitigate risks through early and proactive monitoring.
	→	It was hard to establish all the data we might want to track so for this first pass we narrowed it down to critical areas that were known as potential risks from clinical trials and to major adverse events that would cause concern.
	→	We were able to define our data requirements in the national tool and, once the team had regulatory approval for the study, we set up the study and its purpose on the platform, with purpose-based access requests going out to every GP practice across the country that would be prescribing the drug.
	→	We only receive anonymous data from sites and this feeds into a new national TRE that the regulators have access to and is accessible by our epidemiological and data science team to proactively look for emerging patterns.
AI innovator	→	I work for an AI company, and we have been designing models for use in healthcare.
	→	It has historically been very hard to get data from the market and we usually just get a small number of anonymous records from a few sites to train our models on. We also haven't had any routine way of managing our models and their performance in the market once they are released. This has required a lot of manual work revisiting sites to get more regular data to track performance.
	→	We have our models working with several Trusts and the support burden has become unmanageable for us as an SME.

→ The new national tool has been very helpful. Importantly, we can now get incremental feeds of anonymous data from our client sites, so we can routinely test a model's performance.

→ A new service recently opened up on the national tool where our models can be continually compared to RWD and changes in data that impact our models can be flagged to us automatically. We have just got this working for one of our models across five customer sites and can now see how our models are performing without having to keep asking clients for their data.

→ We have a much better answer now for clients on how we can deal with difficult issues such as context sensitivity and calibration drift in our models.

→ We just had a request from a client to use this mechanism to test our algorithm against one they had built themselves and were operating. We can now do this much more easily and help the client to compare relative performance. This gives us an opportunity to show how our dedicated work in this area has produced a better model.

Palantir

Case Study: National Institutes of Health – National Covid Cohort Collaborative (N3C)

The National Institutes of Health National Center for Advancing Translational Sciences (NIH NCATS), in collaboration with medical researchers around the U.S., is using Palantir Foundry for the National COVID Cohort Collaborative (N3C) Data Enclave.[17] This is a national COVID-19 TRE that is now being leveraged across the world.

The N3C Data Enclave is accelerating research on COVID-19 by giving the medical research community access to the largest shared patient-level data asset of COVID-19 clinical data in the U.S., comprising Electronic Health Record (EHR) data from 75 local medical centres. N3C covers more than 8.5 million patients, of which 2.8 million are COVID-19 positive, and delivers more than 9.7 billion rows of data to serve approved projects.[18]

The challenge

Despite N3C's large data and user scale, NIH NCATS must keep the sensitive EHR data hosted in the Data Enclave safe. It needs to ensure that identifiable information is protected, while providing the access required by researchers for their work, and enabling contributing institutions to maintain control of their own data.

The solution

The N3C Data Enclave uses logical federation to make harmonised, quality-controlled data from local medical centre sites available to researchers in secure shared spaces. When data from each site is ingested into Palantir Foundry, it passes through a harmonisation pipeline to transform it into the OMOP common data model. This harmonisation process is carried out within a workspace that is only accessible to contributor users and administrative users involved in the harmonisation process. Each contributor has its own private space within this environment, where administrative users can work with the NIH NCATS harmonisation team on data quality issues.

Once data has been harmonised, the OMOP tables are combined and made available to a limited group of data quality administrators who can compare data across sites. This information is used to decide which data from which sites should be released to researchers, and to work with contributors to resolve data quality issues that cannot be resolved within the transformation pipeline. Only once data from each contributor has passed rigorous data quality and privacy checks is data moved into a further space in the platform accessible to researchers.

Palantir Foundry's granular access controls promote proper N3C data use by researchers and prevent inadvertent access. The platform backs the provision of purpose-based access controls in the N3C Data Enclave, so that researchers receive access to data according to their specific needs, instead of receiving blanket access at a dataset level.

Upon being granted access to the N3C Data Enclave, researchers must request access to specific N3C datasets at a given tier of information (e.g. synthetic, de-identified Safe Harbour, or HIPAA Limited Dataset data). They make the request and provide justification using an in-platform configured form.

The NCATS Data Access Committee then receives automated notice of each request submission and approves or denies the request in the platform. Palantir Foundry powers true purpose-based access. Should the researcher require this same data for another project, they must submit another access request.

NCATS intends to use privacy-preserving record linkage (PPRL) to link data from its enclave with medical images, omics tools, EHRs and social determinants of health to answer researchers' continuing questions, such as why COVID-19 symptoms persist in some patients.[19] PPRL finds and links records on the same patient across independently maintained data sources using a cryptographic hash value to protect their identity.

Multimodal analytics being implemented now give researchers the ability to look at patient images with their lab results, but some of the data sources NCATS wants to link to the N3C Enclave are maintained by other agencies, like the Centers for Medicare & Medicaid Services. PPRL respects data ownership by temporarily linking datasets in a neutral, high-performance computing area long enough for researchers to complete their work. Duplicated information is also eliminated in the process.

N3C has recently published its insights into synergies between centralised and federated approaches to data quality. It concludes that by combining rapid, continual assessment of data quality with a large volume of multisite data, it is possible to investigate more nuanced scientific questions with the scale and rigour that it requires.[20]

Opportunities

We believe the Framework's model of local data curation and automated, approved data gathering can be leveraged to address a number of the health and social care system's most data-rich challenges, each using the same purpose-based approval mechanisms. This potential for wider use will encourage data controllers to take responsibility for continuously improving the quality and usability of their data assets, a process that will become incrementally quicker and easier over time. Possible applications include:

- → Monitoring the population health impact of new medicines to enable population-based contracting for innovative treatments.
- → Attracting pharmaceutical and medical device companies to the UK as the global standard bearer for the rapid authorisation and adoption of healthcare innovation, enabling them to demonstrate impact at a national scale.
- → Encouraging a fast-growing MedTech industry through providing real time incremental data and an operational environment where MedTech companies can continuously monitor the performance of their innovations against RWD, whilst de-risking these innovations and creating a new safety environment that does not exist today.
- → Enabling NICE to monitor the uptake and impact of its guidance, moving to living guidelines. The ability to monitor RWD data could also support Technology Appraisals, enabling rapid innovations in pharmaceuticals and MedTech that require contingent approvals and further analyses.
- → Enabling the MHRA to adopt data inspection for post-market surveillance and speed up the availability of observational studies for reviewing the safety of regulated technologies.
- → Enabling the NHS to establish the baseline consumption of drugs in order to improve negotiations with the life science industry, and support the move to more novel population-based procurements and, in time, outcomes-based incentives.
- → Enabling research institutes to lower their cost of doing research and spend time on value-added discovery and innovation.
- → Enabling the NHS to speed up the adoption of innovation across the country and offer prompt and fair access for patients.
- → Enabling the UK Health Security Agency to track the outbreak and spread of diseases across the country, without having to provision new infrastructure or commission new data collections each time a public health incident occurs.

This paper has focused exclusively on research and innovation to enable the UK Life Sciences Vision. However, this Framework could equally support wider legitimate and legal uses of data such as healthcare planning and commissioning, population health management, and providing a data clearing house for providers and commissioners for their growing data capture and returns demands, including national registries.

A National Technical Framework offers the opportunity to differentiate the UK in the global healthcare and life sciences market. It could fuel the UK's competitive ambitions following Brexit, and realise the lessons learned during the COVID-19 pandemic.

Most importantly, it would allow care providers, researchers, and citizens to exchange and leverage health data in the most transparent and trusted way, providing a growing data asset that would drive improvements to public health for decades to come.

1 Data saves lives: reshaping health and social care with data, Department of Health and Social Care, July 2021.

2 Life Sciences Vision, HM Government, 2021.

3 The UK trusted and connected Data and Analytics Research Environments (DARE UK) programme, accessed Oct 2021

4 NHS Long Term Plan, NHS, January 2019.

5 Unlocking the value of data: Exploring the role of data intermediaries, Centre for Data Ethics and Innovation, 2021.

6 Trusted Research Environments (TRE): A strategy to build public trust and meet changing health data science needs, UK Health Data Research Alliance, July 2020.

7 Foundations of Fairness: Where next for NHS health data partnerships?, Understanding Patient Data, March 2020.

8 Joining up the dots: driving innovation, research and planning through Trusted Research Environments, NHSX Blog, accessed Oct 2021.

9 Investment to transform access to data to help pioneer new patient treatments, GOV.UK press release, May 2019.

10 Federated queries of clinical data repositories: Scaling to a national network, Journal of Biomedical Informatics, Volume 55, June 2015, pp. 231-236.

11 Accessing secure research data as an accredited researcher: The Five Safes, Office of National Statistics, accessed July 2021.

12 Quote attributed to Sebastian Schneeweiss, Professor of Medicine and Epidemiology at Harvard and Director of the Harvard-Brigham Drug Safety Research Center

13 Sharing of clinical trial data and results reporting practices among large pharmaceutical companies: cross sectional descriptive study and pilot of a tool to improve company practices, BMJ 2019;366:l4217

14 Synergies between Centralized and Federated Approaches to Data Quality: A Report from the National COVID Cohort Collaborative, Journal of the American Medical Informatics Association, Nov 2021.

15 Proposing a common basis for health data access across Europe: 2021 recommendations based on calls to action on health data ecosystems. The European Institute for Innovation through Health Data (i~HD), Accessed Nov 2021.

16 Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. Journal of Medical Ethics. Nov 2019

17 The NICE strategy 2021 to 2026: Dynamic, Collaborative, Excellent, NICE, 2021.

18 National COVID Cohort Collaborative (N3C), statistics accessed August 2021.

19 NIH's COVID-19 data enclave continues to evolve with the virus, FEDSCOOP, accessed August 2021.

20 Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative, Journal of the American Medical Informatics Association, ocab217, Nov 2021.