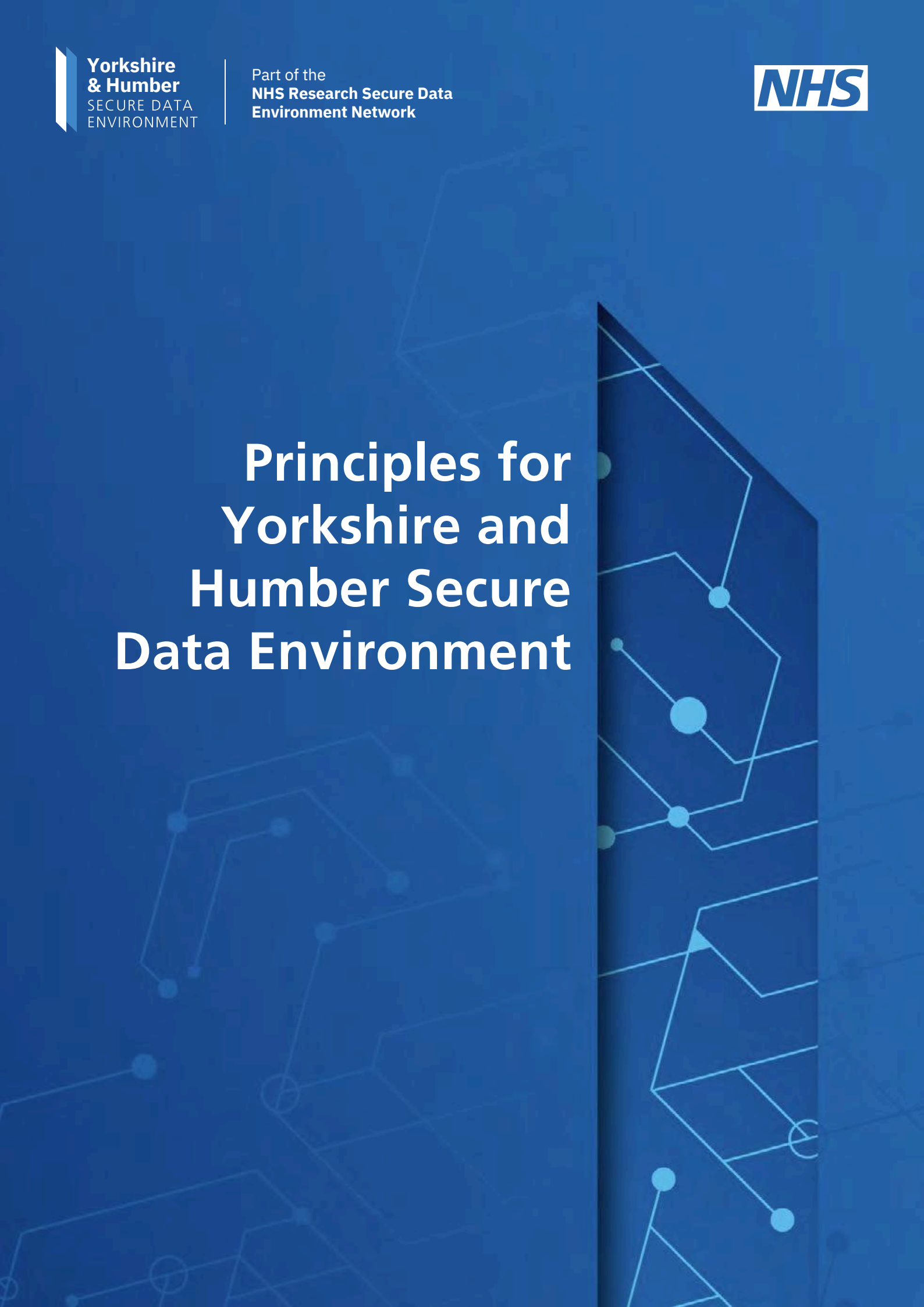
Yorkshire & Humber SECURE DATA ENVIRONMENT

Part of the **NHS Research Secure Data Environment Network**

NHS

Principles for Yorkshire and Humber Secure Data Environment



Yorkshire & Humber SECURE DATA ENVIRONMENT

Part of the **NHS Research Secure Data Environment Network**

NHS

Summary

The Yorkshire and Humber Secure Data Environment (SDE) presents a unified, secure, and transparent platform for managing and utilizing healthcare data, fostering collaboration between local data providers, researchers, and patients. With commitments to transparency, patient-centric privacy measures, and the adoption of standardized data models, the SDE aims to streamline data ingestion, analysis, and egress processes while ensuring compliance with ethical guidelines. By actively engaging with data creators, including General Practice providers, and the public, the SDE strives to improve service delivery, spur innovation, and empower local teams in realizing the benefits of shared data for research and regional healthcare improvements.

The approach taken is deliberately “bottom up”, emphasising patients, data providers, and local control. This is in contrast to many more centralised SDE approaches, but was well-received when it was expressed in our initial Expression of Interest back in 2022. We believe that this approach breaks down many of the barriers encountered by similar projects, and promotes trust by and engagement with providers and the public alike as we focus on this aspect of the data landscape.

Data Providers and Local Involvement

The SDE will encourage local control over data use while also providing support from the central team. This allows for flexibility in how projects are approved and managed, depending on the preferences of the local data provider. It will also work with local data-creating organizations in realizing the benefits of sharing their data and ensuring its appropriate use.

Patient Involvement and Privacy

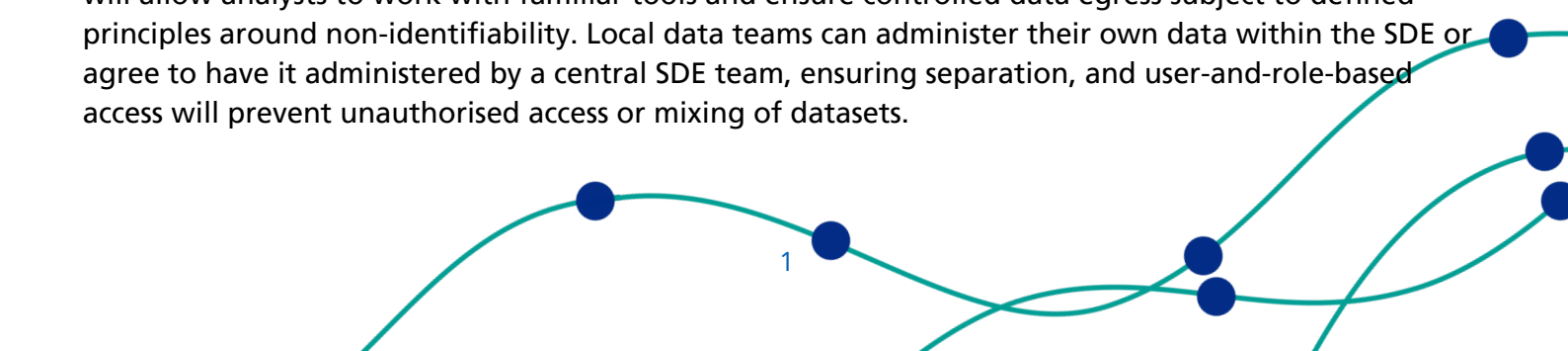
Transparency is a key principle within the SDE, with all projects being publicly listed along with details of their data use. Patient input is also valued, including respecting opt-outs and incorporating patient benefits into project approval processes. Privacy measures include least privilege access to datasets, adherence to standard privacy protocols, and separation between identifiable and pseudonymised data.

Data Federation and Harmonisation

The SDE will encourage the use of a Common Data Model for conceptual and technical harmonization while discouraging data aggregation prior to code being run to ensure that differing processes in routine data do not harm analyses. It aims to present a unified "single front door" for users, streamlining the process regardless of which local provider approves the project or provisions the data, and will actively engage with new providers including General Practice.

Technical Aspects of Data Management

The SDE will support various methods for data ingestion, offer in-situ or remote data analysis to minimise costs and administrative burdens, and enable serverless technologies for efficient analysis. It will allow analysts to work with familiar tools and ensure controlled data egress subject to defined principles around non-identifiability. Local data teams can administer their own data within the SDE or agree to have it administered by a central SDE team, ensuring separation, and user-and-role-based access will prevent unauthorised access or mixing of datasets.



Data Providers and Local Involvement

Local involvement of data teams

For projects involving local data only, it should be possible for a local data provider (if they wish) to approve and set up a project and analysis environment, provision data, and allow analysis to proceed with the SDE central team receiving a report of this activity and not requiring their involvement Equally, with the permission of local teams, the SDE should be able to perform these functions on their behalf. Some may prefer this as a standard model, for others this will only be relevant for projects spanning multiple local data sets.

Local involvement of NHS and other data creators

We recognise that routine data is not the same as research data, in that it is driven by process rather than accuracy – and these processes may differ even between places which are quite physically close, or a single organisation at different times. We therefore will work with local data-creating organisations to ensure that along with the data we have contacts to help researchers understand where the data comes from, so that it is used appropriately. We will encourage researchers to feed back these findings into metadata which can be made available for other researchers working on similar data.
Equally, we must recognise that there is no absolute requirement for organisations to "give up" their data for research. The SDE will support local teams to realise the benefits of sharing their data by encouraging the development of their own questions to improve service delivery and patient care, as well as making use of potential collaborations or data linkages within the SDE.

Onboarding of new data providers

The SDE will actively encourage data providers from around the region to contribute data to the SDE. Recognising that there will be a variety of research and technical literacy, an individualised approach will be taken. If a provider wishes to establish themselves independently, they will be able to make use of suggested approaches, paperwork etc from the SDE to help them and they will be able to retain control. Alternatively elements of this may be handed over to the SDE central or spoke teams, or an existing data provider already using the SDE in an independent manner. Working with Yorkshire and Humber Care Record/Interweave should help ease the burden where data providers already have agreed flows into YHCR and with agreement this can continue on to the SDE. The SDE's principle of transparency will ensure that, even where data providers have delegated some responsibilities, they remain informed of how the data is being used.

Patient Involvement and Privacy

Transparency

The Yorkshire and Humber SDE commits to transparency in how data is being used within the SDE. A list of projects along with the responsible organisation, a description of data being used, and the privacy-preserving techniques being employed will be made public and kept up-to-date. We will encourage code sharing and re-use where possible.

Transparency extends not just to the public whose data we work with, but also the organisations which provide the data. We will maintain engagement with these organisations with reporting of how data is being used and any outputs, or in cases where permissions to approve have not been delegated, liaising with them to allow them control whilst still presenting a single point of contact for researchers.

Patient involvement

We recognise that people are ultimately the owners of their data and data controllers are the custodians, expected to act in the interests of the people. We will respect the opt-out. We will engage with attempts to update the opt-out to allow people a finer control of how their data is used, considering that they may have different feelings on its being used for NHS planning and research vs being used by commercial companies.

Project approval will in part be based on the expected benefits to the people whose data are being used, and researchers encouraged to incorporate plans to benefit the local populace (eg by feeding back results to the local system or ensuring that technology developed can be used locally etc)

Patient/Public engagement will be a core function of the SDE, to keep tabs on the changing landscape of NHS data and its use. The SDE will ensure that the local public are kept informed of its activities, and the benefits that they can see as a result of projects it supports. It will also regularly assess their feelings on how their data is used and incorporate those findings in its methodology for project approval.

Privacy

It is accepted that pseudonymisation is not anonymisation and even pseudonymised patient data can be identified when combined with other datasets. The SDE should therefore operate on the principle of least privilege when it comes to dataset access. A researcher working on multiple projects should not be able to simultaneously access datasets from more than one SDE project (technical solution) and prevented from simultaneously accessing other patient-level data (identified or pseudonymised) by agreements made on sign-up to use the SDE (legal solution). Where possible for a project, technologies along the lines of OpenSAFELY/DataSHIELD and others will be used to avoid researchers needing to have direct access to data.

The SDE will comply with all standard processes to ensure privacy, including ensuring that governance and ethical approvals cover data being used for research purposes, and that research outputs are safe with avoidance of small numbers and adhere to governance processes around other sensitive data.

Pseudonymisation and identification

We will support the use of pseudonymisation-at-source as well as the ingestion of identifiable datasets; identifiable data will be held separately and not made available to researchers. Where re-identification is required by a project (for example, for intervention or trial recruitment) work will occur on pseudonymised data and the reidentified data will land back on the clinical side for use by teams who already have legitimate access to identifiable data (eg the team already caring for a patient or similar)

Data Federation and Harmonisation

Federation

We will encourage the use of a Common Data Model, likely OMOP, to allow for conceptual and technical harmonisation where analyses can be re-run across multiple data sets. However, as we recognise that processes generating data may differ between data sets we will discourage the aggregation of data into "lakes" prior to code being run. Aggregation can occur after the first stage of analysis, so that analysts are able to check that assumptions made about one data set apply to others. We will take an active interest in improving and extending OMOP standards to support the needs of the SDE and the wider NHS, such as by supporting the development of an NHS vocabulary within the system, or vocabularies to support wider non-health datasets.

Single Front Door

Whilst local data teams will be able to retain control and use existing procedures to approve projects, it is important that the SDE is able to present a unified "single front door" to those who wish to use it. SDE proposal forms and processes will be developed, and data providers will be able to use those in addition to or in replacement of their own, and may agree for a central SDE team to approve projects on their behalf, with reporting to the local team. Where this function is not delegated centrally, the central team will coordinate between researchers and the local team to preserve the single front door, and allow local teams to provision requested data into the researcher's workspace. If the data requested is from a single provider, the central team may either put the researchers in touch with the local team for approval and access to be arranged there, or continue to coordinate things centrally.

General Practice

In the landscape of structured NHS data, at present General Practice data is often more complete and usable than standard hospital data forms (although this may change over time as more hospitals develop EPR technology). Additionally, the near-duopoly of EMIS and TPP means that there are few technical standards to accommodate, although processes differing between practices and PCNs can hamper analysis. The SDE will actively engage with these providers, and with General Practice as a group, to find ways of making this data available for use both in its own right and in conjunction with other data such as for supervised AI learning or augmentation of other datasets. Where work such as codelists and data models have been developed, the SDE will work to ensure these are made available across the region and wider network.

Technical Aspects of Data Management

Ingestion

Recognising that different organisations are in a varied state in terms of their technical abilities, we will allow data to be ingested into the SDE with a wide variety of methods rather than attempting to restrict to any individual route. It will be incumbent on the SDE to find a way to support the data provider rather than expecting the data provider to change practices. Routes including SQL server replication, CSV upload, and technologies along the lines of FHIR and Google Cloud Fusion/Azure Data Factory will be available.

Storage

In some cases (eg large multimodal datasets) it may be inappropriate or costly to ingest data directly to store within a new technical infrastructure. The SDE's technical solution should have the capability of analysing data "in situ" by separating the storage from the analytic component. This will also be useful for federated analyses with other SDEs, and potentially for easing the governance side of onboarding new data providers who already have agreements with organisations such as the Yorkshire and Humber Care Record/Interweave.

Analytics

We recognise that research is an inherently "bursty" workload, and particularly that the needs of complex multimodal data analyses can scale up and down by a large degree. The SDE will be able to support these needs, but in a way that controls the financial implications of running powerful compute/GPU by automatically scaling down/turning off servers, and ensuring that research teams can control their billing. We will encourage the use of serverless technologies such as Google Bigquery or Azure Synapse to make analysis more efficient, and will provide a basic shared server as part of standard SDE use which will be suitable for simpler analyses, and preparatory work in support of more complex multimodal analyses.

Equally, we recognise that a variety of tools are in use around the region and that the SDE should enable analysts to work with the tools with which they are most familiar, assuming they have arranged licences for commercial software such as Stata.

Egress

Data egress will be subject to a set of defined principles around non-identifiability, and controlled by the team (local or central SDE) in charge of that project.

Separation

Local data teams will be able to administer their own data within the SDE, and/or agree to have it administered by a central SDE team. This will involve the creation of multiple "admin" workspaces as well as individual user workspaces into which data can be provisioned. Access to the SDE workspace will be in the form of a user-and-role-based-access, ie where a user must select a specific workspace to log into rather than being able to simultaneously access data provisioned for multiple different projects.