Architectural capability for NHS Secure Data Environments

Background

In the Data Saves Lives Strategy, NHS England (NHSE) committed to implementing Secure Data Environments as the default way to access NHS health and social care data for research and analysis. Due to the proliferation of technical infrastructure for storing, managing, and analysing health data across the health system, the NHSE Transformation Directorate has been working to develop policy and guidance to standardise design and deployment approaches for their use. This document is therefore one of a suite of resources which are aimed to support NHS and their partner organisations to deliver secure and interoperable infrastructure.

Purpose of this document

This document intends to set out the core required and preferred architectural capabilities for the use of Secure Data Environments (SDEs) by the NHS, or for the use of NHS data assets by partners, in England. It is focused on the core requirements of SDEs, rather than on specialist deployment for specific data types or specialist analyses.

This capability specification is intended for technical teams who are working within NHS organisations, or organisations managing NHS data (e.g., linking, processing, or analysing), who own and operate SDEs. SDE 'owners' in this context refers to the organisation(s) which set an SDE's design, govern its use and are accountable for the activity carried out within it. The SDE owner will usually be the controller of the data it contains but may be acting as a data processor. There may be multiple organisations working in partnership to deliver an SDE, with roles and responsibilities to meet the capabilities outlined in this document needing to be clearly agreed.

Where something is delineated a '**must**', this means it is a requirement for SDE accreditation - and SDE owners will need to demonstrate that they meet this required capability. A list of these mandatory requirements can be found in Annex 1.

Areas of SDE design covered in this document include:

1. Overarching design
2. Discoverability
3. Standards
4. Cybersecurity
5. Access & approvals
6. Ingress & egress
7. Functionality
8. Data management
9. Auditing & transparency
10. Reproducible analytical pipelines

What is a Secure Data Environment?

Secure Data Environments are data storage, access, and analysis platforms, which facilitate high standards of privacy and security of NHS health and social care data when used for research and analysis.

An SDE could receive data from a number of sources, such as Electronic Health Records or national collections. These systems or applications will not need to be accredited in the same way - although should strive for the same high levels of security and transparency as accredited SDEs.

An example of this is shown in Figure 1.

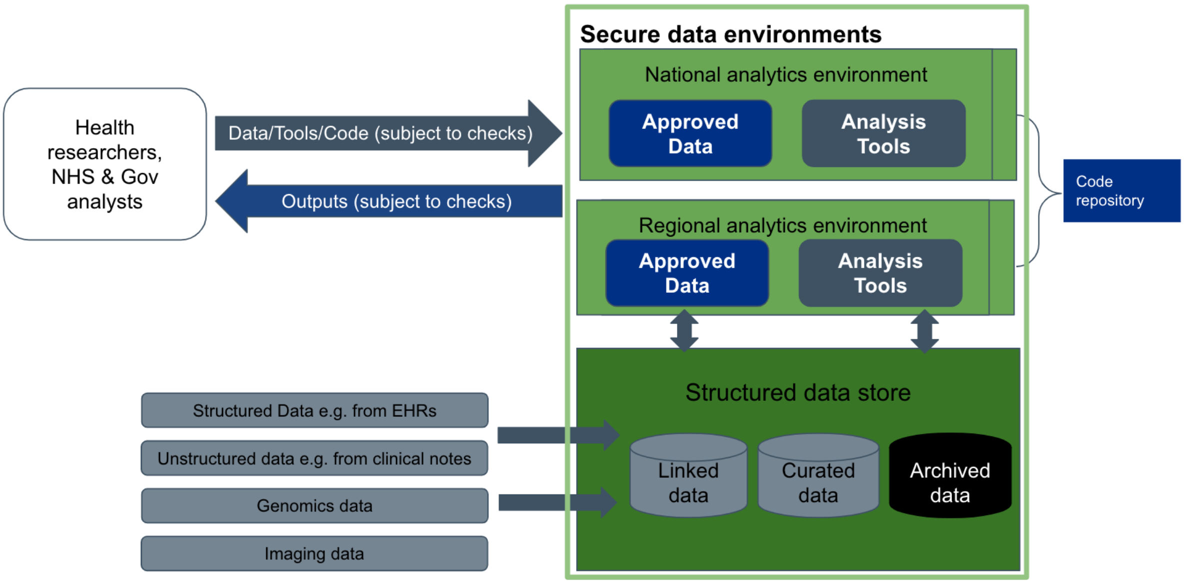

Secure data environments
National analytics environment
Approved Data
Analysis Tools
Regional analytics environment
Approved Data
Analysis Tools
Code repository
Health researchers, NHS & Gov analysts
Data/Tools/Code (subject to checks)
Outputs (subject to checks)
Structured data store
Linked data
Curated data
Archived data
Structured Data e.g. from EHRs
Unstructured data e.g. from clinical notes
Genomics data
Imaging data

Figure 1. Use of SDEs in the English NHS.

While each SDE can use the same infrastructure and entry point for all users, there will be different governance models and attribute-based access controls for varied user groups (e.g., internal analysts, epidemiological researchers, developers etc.) that will determine the functionality available within it.

SDEs are usually not singular units, and will be made up of components including:

- Infrastructure & hosting: the computing power, network usage, and storage that is leveraged.
- Platforms & applications: analytical and data management tools; utilisation of/ communication with external platforms e.g., GitHub.
- Data & data exchange: data storage in databases and warehouses.

These components may be delivered by multiple providers, some of whom might supply more than one aspect of the SDE e.g., a vendor may provide both cloud hosting and an analytical platform.

SDEs should not operate in silos, and if future goals of federated data and analysis are to be achieved, they will need to adhere to a central set of capabilities and design principles. This means that not only do SDEs need to be able to communicate with one another, they need to be able to interoperate with the system wide functionality.

1. Overarching design

There are a number of fundamental design principles and capabilities that all SDEs should adopt in order to facilitate interoperability. SDEs should utilise common software and platforms/tools which have been well tested and demonstrated as robust (with understanding that specialist capability may at times be required). They should draw on open-source offerings as much as possible, with deployment of inappropriate software patterns. They should be built upon open principles and standards, for software interoperability, data, and document formats.

The specific design of each SDE will vary depending on their intended deployment use case (particularly whether they serve to provide access to data or not) but should ensure they have an appropriate compute stack for their intended use, with relevant tooling and compute resources proportional to their users' needs. SDEs should consider providing the ability for users to bring in their own tools as and when deemed suitable and secure, and if there is sufficient resource to support this onboarding.

SDEs should have distinct areas for management of code and data, with security between these (e.g., an air lock or air gap), which uphold principles of data and code portability. There **must** be firewalls established between areas of the SDE and external websites/ repositories. SDEs should support scaling of their storage and compute (where relevant), preferably utilising a cloud agnostic approach (as SDE Owners must meet the guidelines set out by the UK Government's 'cloud first' policy for public sector IT) unless there are legitimate reasons that on premise infrastructure is preferred. SDE owners should refer to the NHS guidance on the use of public cloud services to understand the expectations for secure safeguarding of data when using cloud computing.

SDEs should include services which support handling and managing workload, including monitoring, and observing use across accounts and/or users, ensuring this can be controlled to support cost management. This ensures that there can be transparency in activities that are carried out within it, and that any breaches of privacy and/or security can be traced. There should also be transparency of SDEs security and design approach, with this information included in the data controller/ deploying organisation's Data Protection Impact Assessment and made publicly available for data subjects, in so far as it does not compromise the integrity of the SDE.

Key areas within an SDE

In order to uphold high standards of security and capability, SDEs should ensure they have secure and scalable compartmentalisation within the platform of user workspaces. There should also be distinct areas where code is developed and validated, and where data is provided for analysis. These components can be provided by one or multiple providers, so long as there are appropriate air gaps maintained between them - to ensure that any movement of data, code etc. between them is sufficiently checked and managed, and that no inadvertent access is granted.

There are a number of areas or compartments that SDEs should consider including in their design, to allow for best practice for security, privacy and functionality. Although not mandatory for accreditation, SDE owners are recommended to have separate areas for code development and access to data for analysis.

These areas include (for all SDEs)

- An airlock: a secure area, isolated from other areas of the SDE and network connections, where any data, code, or tools to be brought into an SDE are first uploaded for review (drawing on automated and human checks as required). Similarly, all outputs and codes being removed from an SDE should be uploaded to a secure platform prior to release, or when moving between data and analytical layers of an SDE. This is separate from the portal by which users are approved for and onboarded into SDEs. [GEL case study]

These areas include (for SDEs with analytical functionality):

- Users should have a dedicated, secure compartment in which to conduct their particular project - this should include a compute or landing area where users can kick off training and testing jobs and monitor their progress.

- A code development area: SDEs should provide an area where code can be, at a minimum, validated (and assembled if appropriate) prior to being run against data. This should allow connections to external repositories, to allow and encourage collaborative research from multiple institutions, but with appropriate firewall provisions. For example, firewall considerations for a GitHub account can be configured to block write back functionality. This area should act as a development area where users are able to create learning projects, upload their algorithms, define a data query, and tweak its parameters.

- A data analysis/ access area: This is the only part of the SDE where data should be accessed by users (with virtualisation options an operational choice). Data from multiple sources may need to be loaded into staging areas for validation, transformation, and normalisation. GPUs may be offered as part of the compute service provided, depending on the deployment context of the SDE.

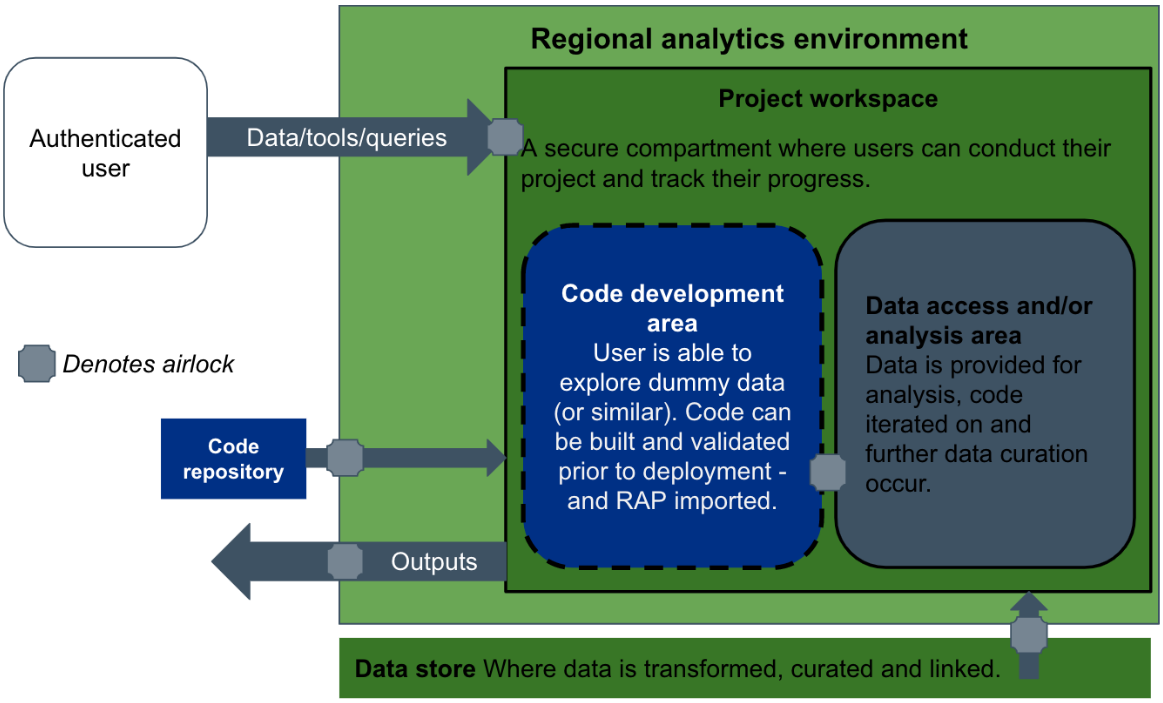

Regional analytics environment
Authenticated user
Data/tools/queries
Project workspace
A secure compartment where users can conduct their project and track their progress.
Denotes airlock
Code development area
User is able to explore dummy data (or similar). Code can be built and validated prior to deployment - and RAP imported.
Data access and/or analysis area
Data is provided for analysis, code iterated on and further data curation occur.
Code repository
Outputs
Data store Where data is transformed, curated and linked.

Figure 2: Key areas to be included in an analytical SDE.

[NW London Discover Now case study]

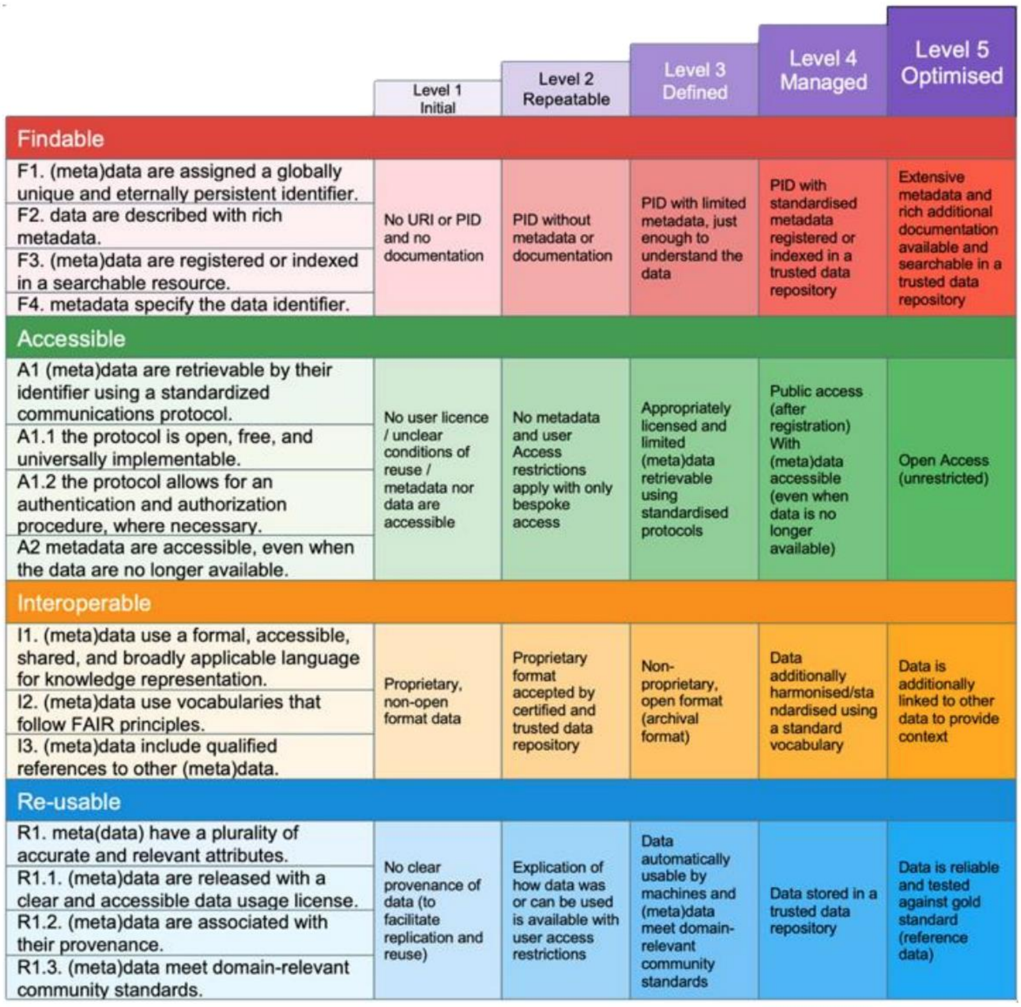
2. Discoverability

Potential users should be able to understand the high-level contents of the data available, how it has been used previously and relevant details of its lineage - ensuring they can make comprehensive and appropriate access requests.

- All SDEs which store data **must** have a discoverability interface, such as a metadata catalogue. This should provide a standardised self-service platform for organisations to browse datasets and associated metadata and data models to determine which information they require for their project. Metadata should also include the analytical environment(s) which can be leveraged to access data assets, including whether they are part of a trusted federated network.

- Metadata **must** be kept up to date on activities carried out with the data it represents, as well any issues or quality markers. Its content should be available in human and machine-readable formats, to support monitoring and aggregation. Metadata should be kept up to date on activities carried out with the data it represents, as well any issues, curation information or quality markers.

The depth of detail available in metadata should be proportionate to the underlying dataset, and follow the Findable, Accessible, Interoperable and Reusable (FAIR) principles as far as possible. Figure 3 shows how SDE owners can manage their metadata lifecycle to provide quality insights for efficient data sharing.



	Level 1 Initial	Level 2 Repeatable	Level 3 Defined	Level 4 Managed	Level 5 Optimised
Findable					
F1. (meta)data are assigned a globally unique and eternally persistent identifier. F2. data are described with rich metadata. F3. (meta)data are registered or indexed in a searchable resource. F4. metadata specify the data identifier.	No URI or PID and no documentation	PID without metadata or documentation	PID with limited metadata, just enough to understand the data	PID with standardised metadata registered or indexed in a trusted data repository	Extensive metadata and rich additional documentation available and searchable in a trusted data repository
Accessible					
A1 (meta)data are retrievable by their identifier using a standardized communications protocol. A1.1 the protocol is open, free, and universally implementable. A1.2 the protocol allows for an authentication and authorization procedure, where necessary. A2 metadata are accessible, even when the data are no longer available.	No user licence / unclear conditions of reuse / metadata nor data are accessible	No metadata and user Access restrictions apply with only bespoke access	Appropriately licensed and limited (meta)data retrievable using standardised protocols	Public access (after registration) With (meta)data accessible (even when data is no longer available)	Open Access (unrestricted)
Interoperable					
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta)data use vocabularies that follow FAIR principles. I3. (meta)data include qualified references to other (meta)data.	Proprietary, non-open format data	Proprietary format accepted by certified and trusted data repository	Non-proprietary, open format (archival format)	Data additionally harmonised/standardised using a standard vocabulary	Data is additionally linked to other data to provide context
Re-usable					
R1. meta(data) have a plurality of accurate and relevant attributes. R1.1. (meta)data are released with a clear and accessible data usage license. R1.2. (meta)data are associated with their provenance. R1.3. (meta)data meet domain-relevant community standards.	No clear provenance of data (to facilitate replication and reuse)	Explication of how data was or can be used is available with user access restrictions	Data automatically usable by machines and (meta)data meet domain-relevant community standards	Data stored in a trusted data repository	Data is reliable and tested against gold standard (reference data)

Figure 3: Metadata capability maturity model, according to the FAIR principles.

SDEs should consider providing a mechanism for exploring data prior to undergoing the full access application process, to support potential users in their discovery work.

[HDRUK Innovation Gateway case study]

3. Standards

The application of standards to SDEs is dependent on the functionality it is intended to provide, with standards needing to be applied to all or some of an SDE's components. SDEs should, in so far as is possible, use standards which are open, replicable, and free to use. The use of open standards supports interoperability and flexibility for both data and software, as well as avoiding vendor lock in and improving sustainability. Additionally, open standards allow users to copy, distribute and use technology freely or at low cost.

The government's Open Standards principles lays out the best practice approach to open standards use - and serves as a strong basis which the NHS should build on. This states that open standards which are used by government should be:

- Well documented, publicly available, and free to use to provide fair access.
- From a collaboration between all interested parties, not just individual suppliers.
- Backed by a transparent and published decision-making process that is reviewed by subject matter experts.
- Developed following transparent and published feedback and ratification process to ensure quality.
- Supported by the market to demonstrate the independence of platforms, applications, and vendors.
- Released for use with a royalty free licence which is irrevocable unless there is a breach of conditions.
- Compatible with both open source and proprietary licensed solutions.

SDE owners should refer to the UK Government Technology Code of Practice, which provides guidance on designing and procuring technology in key areas including accessibility, security and integration.

The standards specifically detailed in the following sections are to be included in the SDE accreditation framework, with a related timeline of expected compliance.

4. Cybersecurity

Although maintaining a high standard of cybersecurity goes beyond technical measures to include integration of manual human touch points throughout the process, there are a number of actions which SDE owners should take to ensure the highest standards are upheld.

Networked access

- SDEs should include a ‘protection layer’ e.g., virtual private cloud, to prevent nefarious or inappropriate access.
- SDEs should minimise external connections across secure environments and use a central server where possible.
- Networks should be separated by an airlock (see section on ingress & egress) or gap, with segment separation of data according to de-identification risk.
- The Transport Layer Security protocol should be applied to communications across and between networks.
- If access to external repositories is permitted in the SDE design, these should be appropriately firewalled and only be accessible from the sandbox or code provisioning zone (or equivalents) to ensure that data is not inappropriately accessed through these.
- The route of access to an SDE is an operational decision for the SDE owner - e.g. can be a VDI, web browser etc. - so long as it is demonstrably secure.

Standards

There are a number of specific cybersecurity standards that SDEs **must** adhere to:

- ISO 27001 - Information Security Management.
- NHS Digital cloud security good practice guide (where applicable).
- NHS Digital Data Security and Protection Toolkit compliant.

- Cyber Essentials Plus: Supplier has been independently assessed and verified by a Government approved external body that it meets the Cyber Essentials implementation profile [BIS/14/696].
- The Security of Network and Information Systems Regulations 2018 (where applicable).
- Data Centre Alliance Class 3 Facility European Code of Conduct (EUCOC): Compliant.

Encryption

- Data stored and utilised in SDEs **must** be encrypted while at rest and in transit. As an example approach, the National Cyber Security Centre provides useful information on encryption and how to protect your data in transit.

- Data controllers **must** ensure encryption of identity linkage keys. These should be held by the data controller, although there may be some exceptions where a trusted broker may be used - the appropriateness of this will need to be assessed during the accreditation process.

Privacy Enhancing Technologies

These are tools and processes which serve to protect the privacy and confidentiality of sensitive data, and include de-identification techniques and encryption approaches, as well as the use of SDEs more generally. The PETs that an SDE should deploy will depend on the users and use cases it intends to support. The emergent nature of many of these technologies means there isn’t current consensus on how they should be applied in the health data context. SDEs should therefore be built on flexible infrastructure, which will allow for responsiveness to the evolving work in this field.

The Centre for Data Ethics and Innovation have developed an adoption guide for PETs - an interactive tool designed to aid decision-making around the use of PETs in data-driven projects. Primarily aimed at technical architects and product owners, it provides guidance on what PETs should be used under a variety of different usage contexts, as well as definitions of each.

Testing

- SDEs **must** undergo penetration and security testing by a third-party authority prior to, and at reasonable intervals during, deployment (e.g., annually). Security testing should include threat modelling by internal security teams, as well as independent auditing of infrastructure and practice.

- Security dashboards should be made available for review with information on security operations monitoring, and any breaches in standards reported to the accreditation body.

5. Access & approvals

Robust identity management and access control are key to upholding high cybersecurity standards and **must** support multi factor authentication of users.

- SDEs which intend to provide access to or allow analysis of data, such as TREs and ODEs, **must** interoperate with any user authentication portal via the SAML or OIDC protocol and avoid non-standard technologies if possible.

- SDEs **must** be able to support decoupling user authentication (their ‘passport’) and approval for access (their ‘visa’). This will provide capabilities for approved user access to be granted and relevant parties to be automatically informed.

- SDE owners should consider the legal mechanisms, e.g., asking external researchers to sign a contract, which can be used to enforce user behaviours - with use of electronic signing and recording of permissions/ expectations bringing the ability to automate their management. In addition to contractual documents there should be a Security Operating Procedures (SecOps) document or equivalent that requires individual users to sign up to a code of practice for using the SDE and can be used to revoke/sanction anyone in breach of those terms.

Attribute based access controls

- SDEs should be able to verify individuals’ identity and eligibility, with functionality granted based on attributes (user attributes, resource attributes, environment attributes etc.), rather than their role/ job title. This will allow SDE owners to assess users’ skills and organisational relationships to determine what degree of functionality they are able to access.

- There **must** be recording of individual user preferences and approvals within the TRE, and the ability to revoke or suspend access when required.

Purpose based access controls

Purpose based access controls consider the users' specific request when determining access but is distinct from access requests to data from custodians.

- Purpose based access controls should assess:
    - The data that has been approved for use (and whether it will be viewed or manipulated) and the risk of data subjects being re-identified.
    - The use of PETs, and expectations or requirements for their use.
    - The analytical functionality required, and whether users will expect to be able to view and/or manipulate the data.

There are no set risk thresholds for re-identification of data, or current standardised approach to carrying out a privacy risk assessment (although work is ongoing in this area across a number of academic partners).

6. Ingress & Egress

Secure ingress and egress processes will require the use of an airlock mechanism. How such an airlock is designed is at the discretion of the SDE owner, with more rigorous checks and manual rather than automated review required for projects deemed 'higher risk' (e.g., due to the nature of the data or intended outputs).

Data

It is assumed that data linkage will predominantly take place in SDEs for data storage and warehousing (the 'data layer'), rather than in those used to provide access to it (the 'analytical layer'). This is to ensure that encryption keys are always held by the data controller, or a trusted broker [Case study: UCL & Moorfields], and to reduce the risk of inappropriate disclosure. There are, however, instances where aggregated data which doesn't require personal identifiers for matching records may be linked within an analytical SDE e.g., analysis of rates of disease against air quality levels, which are linked per geographical area.

Data **must** be ingressed through an air lock to the data provisioning zone (or equivalent secure area within a users workspace), directly from the ‘data layer’. It should be checked to ensure that:

- The user takes responsibility for the content.
- It is the correct data for the intended workspace/ user(s).
- It is in the appropriate format for analysis (e.g., right data type(s), language, size).
- It is sufficiently de-identified when entering the analysis layer (in accordance with project approvals).
- There are no viruses or malware present.

Where required, data pipelines can be approved through an air lock before deployment and then run continuously, with re-review through the airlock only required if there is a change to the analytical code or underlying data. This will ensure that any ongoing feeds of data or insights will be assured as secure, while not creating a burdensome process for SDE owners.

Code

Any analytical code entering the SDE, whether from within the code provisioning zone (or equivalent), from an external repository or from another SDE via federated learning processes, needs to be screened to ensure:

- That code is in an appropriate format and able to run against the data points it is to be used on.
- It is in line with the project documentation that has been approved.
- It is not malicious, e.g., going to cause harm, inadvertently contains personally identifiable data, or attempt to remove or re-identify data.
- There are no viruses or malware present.

This **must** be done utilising an air lock mechanism.

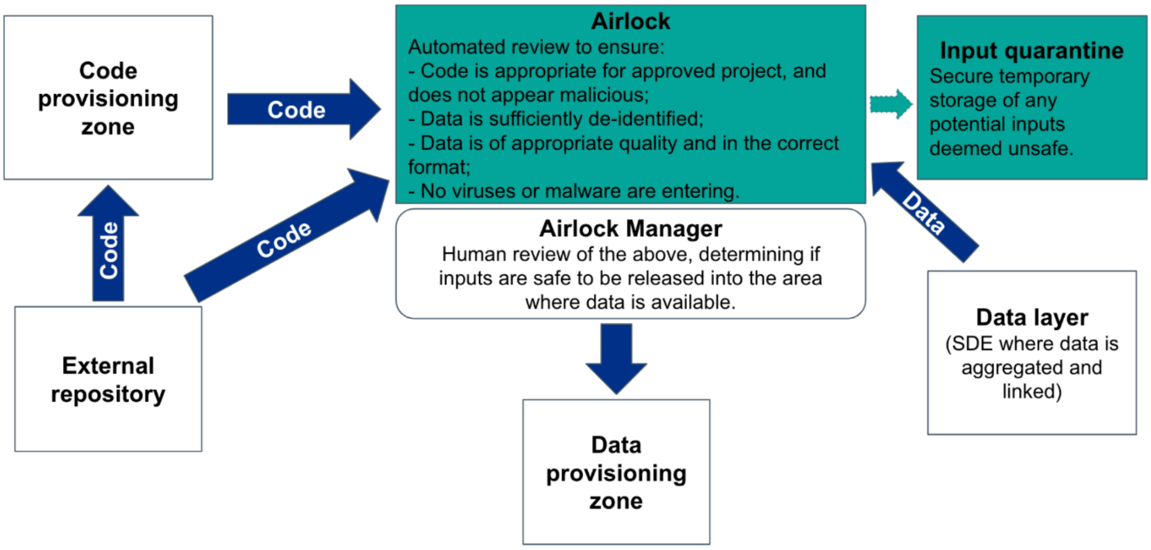

Code provisioning zone
Code
Airlock
Automated review to ensure:
- Code is appropriate for approved project, and does not appear malicious;
- Data is sufficiently de-identified;
- Data is of appropriate quality and in the correct format;
- No viruses or malware are entering.
Input quarantine
Secure temporary storage of any potential inputs deemed unsafe.
Code
Code
Data
Airlock Manager
Human review of the above, determining if inputs are safe to be released into the area where data is available.
External repository
Data layer
(SDE where data is aggregated and linked)
Data provisioning zone

Figure 4: Example of the ingress process for an analytical SDE.

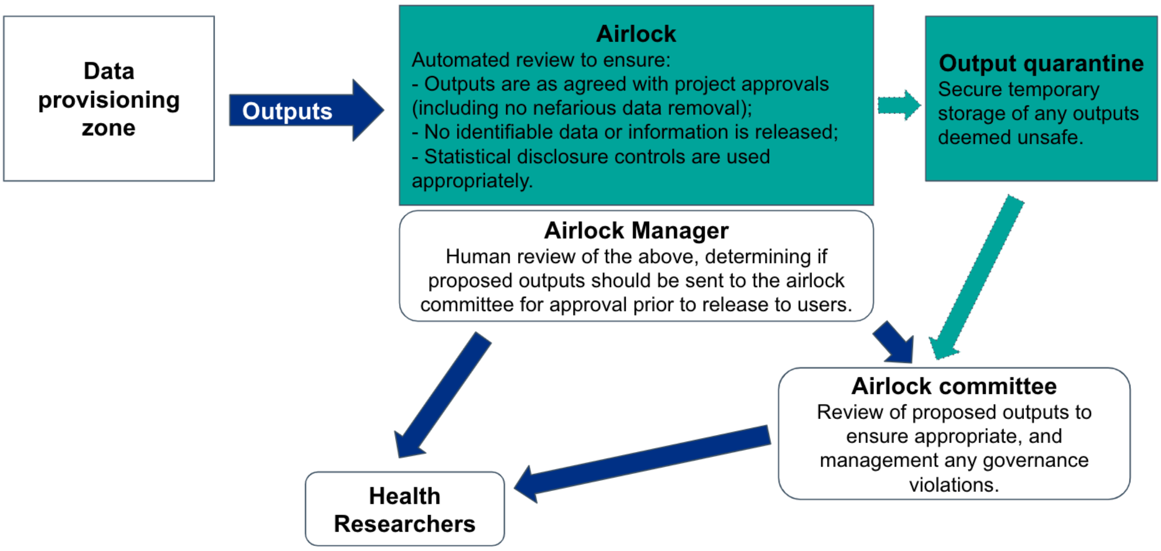
Outputs

Any outputs (e.g., data insights, analytical code and updated algorithmic models etc.) from analysis should be egressed safely, with screening to ensure:

- They are in line with the project documentation that has been approved.
- There is no inappropriate removal of data.
- That the content is in line with the agreed project approval, including appropriate application of statistical disclosure controls.
- That the user takes responsibility for the egressed content and understands any contractual restrictions on publication and re-use (where applicable).

Again, an air lock **must** be used to review outputs prior to release. As with data feeds, analytical pipelines can be approved once through an airlock process before continuous deployment, with re-review only required if there is a change to the code or underlying data. For code under systematic, iterative development, SDEs will also need an inwards only ‘code import pipeline’ e.g. via a proxied git service.

[Case study: PHE Fingertips]



Data provisioning zone
Outputs
Airlock
Automated review to ensure:
- Outputs are as agreed with project approvals (including no nefarious data removal);
- No identifiable data or information is released;
- Statistical disclosure controls are used appropriately.
Output quarantine
Secure temporary storage of any outputs deemed unsafe.
Airlock Manager
Human review of the above, determining if proposed outputs should be sent to the airlock committee for approval prior to release to users.
Airlock committee
Review of proposed outputs to ensure appropriate, and management any governance violations.
Health Researchers

Figure 5: Example of the egress process for an analytical SDE.

If there is any inappropriate user activity suspected, then there should be functionality with the SDE to (1) quarantine any data, code, outputs or tools, and to (2) suspend a users access until human review (by air lock manager, committee etc.) has taken place.

7. Functionality

SDEs which allow access to and/or analysis of data should provide 'layers' of functionality mapped to their access and approvals framework. They do not need to provide every layer of functionality, but should ensure that these are cumulative i.e., that if the 'highest' level of functionality is provided then the lowest is available too.

SDEs should support running and management of packages, clusters and workspaces as required. These should persist for the duration of the project, following which they may be archived (e.g., if required for evidence or regulatory purposes) within the SDE or removed. They should also be able to allocate appropriate compute capacity, recording and monitoring usage against allocations (according to data sharing agreements, commercial agreement etc.) where relevant.

The functionality and tooling provided should be proportional to the user (their experience, training and intended purpose/outputs), data type and risk of data re-identification. SDEs should strive to provide up to date tooling for users, and support common languages e.g., Python. Each SDE can set its own layers of functionality according to its need, but generally speaking to manipulate data users will have to undergo more rigorous governance and demonstrate increased levels of competency compared to viewing it or simply querying it.

All levels of SDE functionality should:

- Provide secure workspaces for users, which prevent exfiltration of data but with archiving functionality available.
- Support validation of queries prior to being run on data, to ensure they are functional.
- Prevent users from copying and pasting data or code from within the SDE.
- Allow users to version control their analysis and, where relevant, data in use.

Users

Linked to attributes as detailed in access & approvals section.

- Individuals should be able to provide evidence that they meet the relevant criteria needed to access the functionality layer they require. If this is not met then the SDE should be able to flag what attributes they require, and the user will need to update their credentials.
- Users' activity within the SDE should be recorded, with higher risk layers of functionality having increased levels of both detail recorded and scrutiny of activity.

Data type

SDEs need to be able to implement variable levels of controls, according to the type of data that will be provided for analysis.

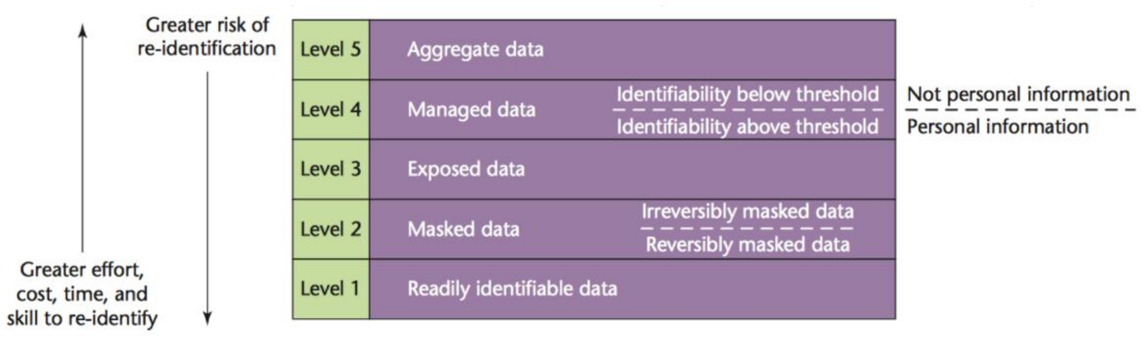
- The type of data being analysed shouldn't necessarily affect the functionality granted for its access or analysis - with the risk of its re-identification the key consideration for the degree of access and functionality given to a user.
- Personal data is personal data whether in image or text or another format, and as such the re-identification risk assessment carried out should focus on the data's variables and parameters, as well as the users ability to carry out identification activities. Re-identification risk also has to account for the environment the data will be available in, such as the users, security processes and oversight structure.

We do know, however, that patients and the public have said in previous research that they view certain data types - such as genomics or mental health data - as more sensitive. Design of SDE functionality layers (and associated approvals processes) should therefore take into account the outcome of any patient and public engagement work carried out with the population whose data it pertains to.

Data reidentification

- Calculating the risk of re-identification should consider the likelihood of making a distinguishable line between ‘could’ and ‘will’ and the impact that it will have so that the risk is managed but not eradicated.
- There may be instances where potentially personally identifiable data can be made available, as appropriate consent is in place from the data subjects.

Generally speaking, data which is richer and/or at individual level is at higher risk of re-identification. The spectrum of data against likelihood of re-identification is shown in Figure 6.


Greater risk of re-identification
Greater effort, cost, time, and skill to re-identify
Not personal information
Personal information

Level 5	Aggregate data	
Level 4	Managed data	Identifiability below threshold Identifiability above threshold
Level 3	Exposed data	
Level 2	Masked data	Irreversibly masked data Reversibly masked data
Level 1	Readily identifiable data	

Figure 6: Showing the goal of de-identification and risk associated with each level, from “Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification”.

The degree to which aggregation can prevent re-identification depends on the sample size and what other data might be linked that could support identification. Similarly, the degree of data masking undertaken also impacts how much it reduces re-identification risk, e.g., just of direct identifiers or of quasi-identifiers (e.g., date of birth to age) too.

Validation

- Validation of analytical models, ensuring they will be able to run against the intended dataset, should be carried out utilising a subset of the dataset (with caution over potential risk of misuse or re-identification), dummy data or synthetic data.

Dummy data is data which retains the original format but doesn’t replicate underlying patterns. Synthetic data definitions follows the ONS nomenclature of synthetic datasets being either 'Synthetic structural' (preserving only the format and variable types) and 'synthetic valid' (as above and also ensuring that a combination of variables per one record is plausible according to the original data). Anything else would be considered 'synthetically augmented' and so would need to be assessed for disclosure risk.

Currently there is no technology which can generate appropriate dummy or synthetic data for key NHS datasets (e.g., SUS) or for relational datasets (e.g., activity crossing primary and secondary care) that would allow code to be meaningfully validated - and therefore alternative approaches to validation should be explored.

8. Data management

Consistency in how data within SDEs is managed is necessary to ensure that it is accessible and of sufficient quality, and that the privacy rights of data subjects are upheld. Although there are no specific data standards being mandated for accreditation, SDE owners should consider how they are able to accommodate commonly used standards - including SNOMED, OMOP, FHIR and DICOM formats.

Data de-identification

- The aim of de-identification is to reduce the risk of re-identification to an acceptable level while retaining as much data utility as possible (functional). This is in contrast to removing the risk altogether which would inevitably deem the data unusable (absolute). UKAN’s anonymisation decision making framework is a practical guide that gives more operational advice than the ICO’s Anonymisation Code of Practice and can be used as a guide to achieving GDPR-compliant anonymisation.
- Pseudonymisation algorithms should be built on open standards. The ICO is working to update its guidance on approaches to pseudonymisation.

- SDEs should include automated removal of individuals from data assets shared via section 251 approvals who have opted out of having their data included in research and planning. They should be able to respond to changes in patient preferences for new and per analysis projects, rather than historic or published, and remove data in a timely manner.

Data storage & systems

- Any data assets which are created or used within an SDE should be considered for archiving - particularly if they may be required for regulatory or replication purposes in the future. This includes the ground truth utilised for any AI training projects. These should be retained in an SDE which provides the 'data layer', managed by the data controller (or one of them in multi controller arrangements).
- SDEs should use a data management system appropriate to the data held within the SDE (e.g., structured, unstructured, or multimodal data). Data should not be copied and should not persist within an analytical environment.

Data flow & provision

- SDEs **must** not allow sharing of data that is not sufficiently anonymised for the intended user, or to unapproved users.
- Minimisation should be applied to any data being utilised, whether it is viewed/ manipulated or not, so that only approved variables are included.
- ETL technology utilised by SDEs should be scalable, flexible and have reasonable fault tolerance.
- SDEs should have the ability to receive data on demand.
- SDEs should be enabled to allow data to be transformed into research ready formats such as OMOP.

Data provenance

- Any datasets curated for analysis need to be recorded, and should be watermarked or similar, where such technology is available for the data type in question. This will allow for its integrity to be monitored.

- As data is transformed during the course of a project, it should be version controlled and time stamped - with SDEs providing archiving ability to securely store data when required (ie as agreed within project approvals), either as a dataset or within a user workspace. This archived data (and any associated analysis where relevant) should only be accessed for evaluation and quality assurance processes by approved SDE or regulatory staff.

- The origin of data/ a dataset needs to be recorded in its metadata, along with any curation and transformation it has gone through. Users may request to apply their own curation code to data, but the underlying version should not be irrevocably changed. A copy of any curation code should also be retained for re-use on other datasets, to support consistency in further curation activities and standardisation of data formatting. Such code should be made publicly available where possible.

9. Auditing & transparency

Recording of activity within SDEs for future scrutiny is a key element in building trust in both data security and output validity, as well as fostering collaboration.

Auditing

- SDEs **must** be able to audit the activity that takes place within them, with a record kept of queries that have been run (proportional to the data and type of analyses), alongside the relevant user(s) and their purpose. These should be retained for an appropriate time frame, as deemed necessary by a relevant regulatory or statutory agency's guidance e.g., MHRA.

Transparency

- Code and outputs should be made available for peer review (subject to reasonable justifications for commercial and academic sensitivities), through the use of collaboration software, open repositories and Reproducible Analytical Pipeline processes - with analytical SDEs able to facilitate committing of code and workspaces to repositories.

- Auditing information **must** be stored securely, and made available for review e.g., in a data use register. TREs should support central aggregation of auditing information by storing information in a machine readable format and be interoperable with national infrastructure. Where reasonable justifications against public sharing are met, these should only be accessible by approved staff.

10. Reproducible analytical pipelines

Reproducible analytical pipelines (RAP) are automated statistical and analytical processes, and their use can improve analytical efficiency and quality. They have a proven track record across government, and the Goldacre review on the use of health data for research and analysis has centred their use as key to realising modern, open, collaborative approaches to data science across the health system.

- SDE owners should ensure they are able to support analysts to leverage RAP, with these the default process for SDE users, albeit with reasonable justifications due to commercial or academic sensitivities. SDEs should therefore supply the tools analysts need to adopt RAP principles and help them to re-use each other's work.

  The government RAP strategy outlines the following tools as minimum requirements for RAP:
  - version control software, e.g., git.
  - open-source programming languages and flexibility to add more – Python, R, Julia, JavaScript, C++, Java/Scala and so on.
  - package and environment managers for each of the available languages.
  - packages and libraries for open-source programming languages, either through direct access to well-known libraries, for example, npm, PyPI, CRAN, or through a proxy repository system, for example, Artifactory.
  - individual storage, for example, home directory.
  - shared storage, for example, s3, cloud storage, with fine-grained access control, accessible programmatically.
  - integrated development environments suitable for the available languages – RStudio for R, Visual Studio Code for Python and so on.

- A repository for the code and artefacts of RAP pipelines should be created and maintained by each SDE, with a need for SDE owners to determine what will be public facing, and what will be available only to approved users (e.g., due to the nature of the analyses). This should be Git based (although this is not mandatory) with users allowed to pull in from it, but not push back out from within the SDE. Responsibility for ensuring quality and safety of code will need to be considered within project management.

Implementation

The transition to using Secure Data Environments in the use of NHS health and care data is a positive step forward. However, it is a complex and rapidly developing field with a range of maturity levels across the country. Therefore, as we gather further insights through research, investments, and accreditation, further detailed technical documentation will be developed - building on these foundational capabilities.

We will continue working with a wide rake of stakeholders to develop and publish information about plans and timescales for transition and implementation.

Annex 1 - The mandatory requirements on SDEs to meet accreditation

In order to become an NHS accredited SDE, these environments must be able to demonstrate they meet the following requirements:

Discoverability

- Have a discoverability interface, such as a metadata catalogue.
- Keep metadata up to date on activities carried out with the data it represents, as well any issues or quality markers.

Cybersecurity

- Adherence to:
  - ISO 27001 - Information Security Management.
  - NHS Digital cloud security good practice guide (where applicable).
  - NHS Digital Data Security and Protection Toolkit compliant.
  - Cyber Essentials Plus: Supplier has been independently assessed and verified by a Government approved external body that it meets the Cyber Essentials implementation profile [BIS/14/696].
  - The Security of Network and Information Systems Regulations 2018 (where applicable).
  - Data Centre Alliance Class 3 Facility European Code of Conduct (EUCOC): Compliant.
- Firewalls established between areas of the SDE and external websites/ repositories.

- Data stored and utilised in SDEs encrypted while at rest and in transit.
- Encryption of identity linkage keys by data controllers.
- Performance of penetration and security testing by a third-party authority prior to, and at reasonable intervals during, deployment.

Access and approvals

- Carry out multi factor authentication of users.
- Interoperable with the OIDC and SAML protocols to facilitate users to access the data they hold.
- Able to support decoupling user authentication (their ‘passport’) and approval for access (their ‘visa’).
- Record individual user preferences and approvals, with the ability to revoke or suspend access when required.

Ingress & Egress

- Data and code to be ingressed through an air lock.
- Outputs to be egressed through an air lock.

Data management

- Not allow sharing of data that is not sufficiently anonymised for the intended user, or to unapproved users.

Auditing & transparency

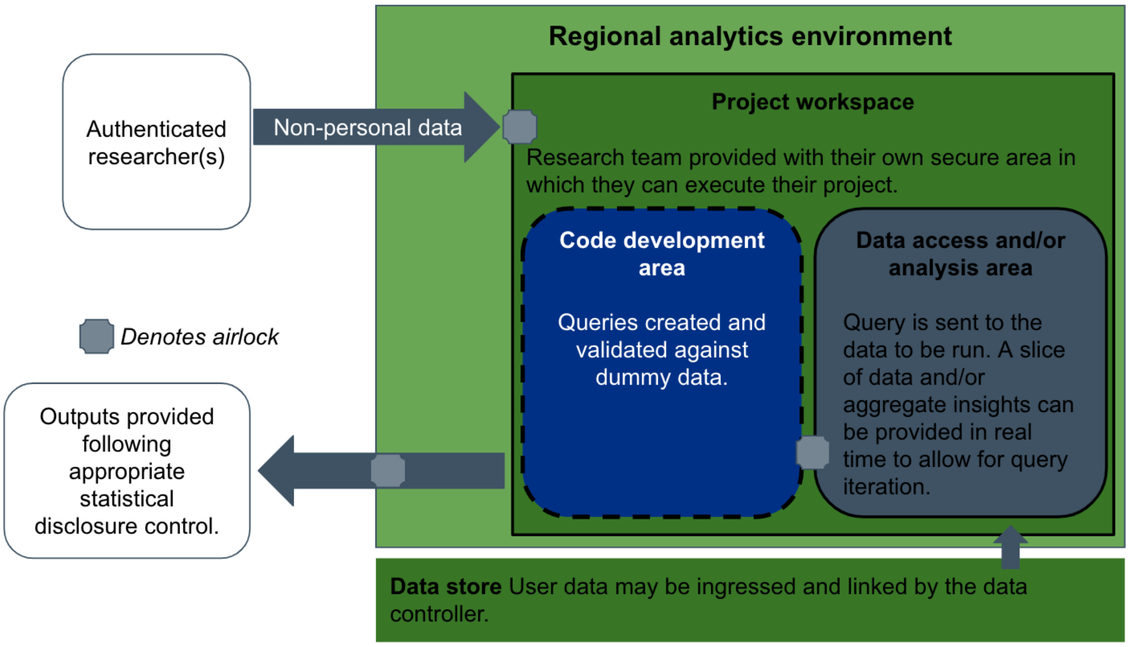
- Able to audit the activity that takes place within them, with a record kept of queries that have been run (proportional to the data and type of analyses), alongside the relevant user(s) and their purpose.
- Auditing information stored securely, and made available for review e.g. in data use registers.

Annex 2 - User personas

The granular design approach and controls implemented should be dependent on the user and their intended purpose for accessing data through the SDE.

Examples of SDE users and how deployment may need to accommodate their needs include:

1. Epidemiological Research



Regional analytics environment

Project workspace

Research team provided with their own secure area in which they can execute their project.

Code development area

Queries created and validated against dummy data.

Data access and/or analysis area

Query is sent to the data to be run. A slice of data and/or aggregate insights can be provided in real time to allow for query iteration.

Authenticated researcher(s)

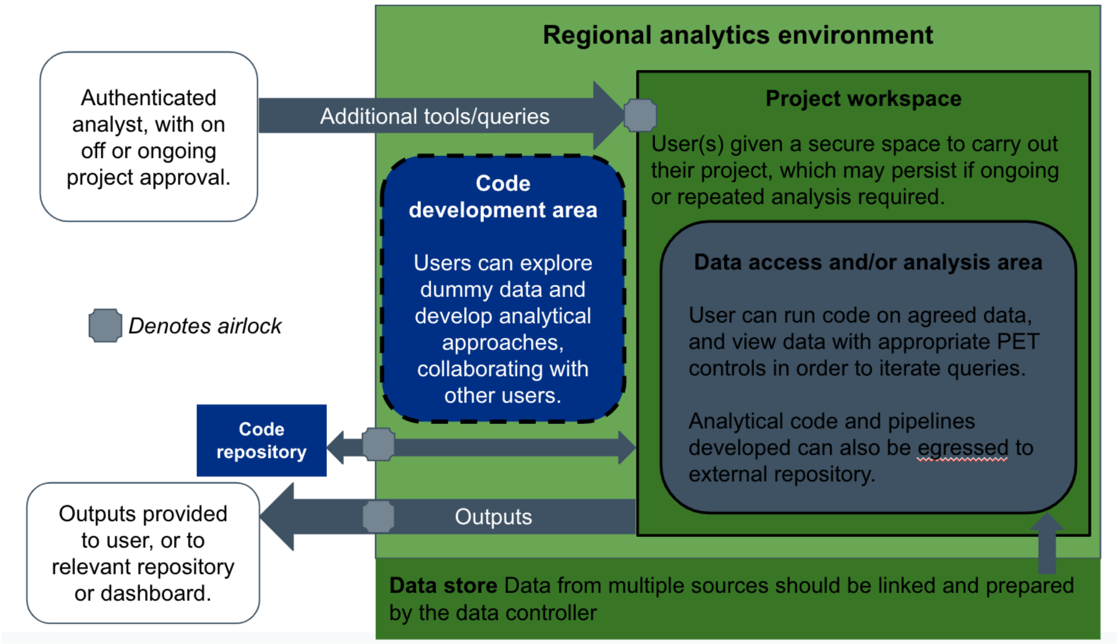
Non-personal data

Denotes airlock

Outputs provided following appropriate statistical disclosure control.

Data store User data may be ingressed and linked by the data controller.

2. NHS analysts



Regional analytics environment

Project workspace

User(s) given a secure space to carry out their project, which may persist if ongoing or repeated analysis required.

Code development area

Users can explore dummy data and develop analytical approaches, collaborating with other users.

Data access and/or analysis area

User can run code on agreed data, and view data with appropriate PET controls in order to iterate queries.

Analytical code and pipelines developed can also be egressed to external repository.

Authenticated analyst, with on off or ongoing project approval.

Additional tools/queries

Denotes airlock

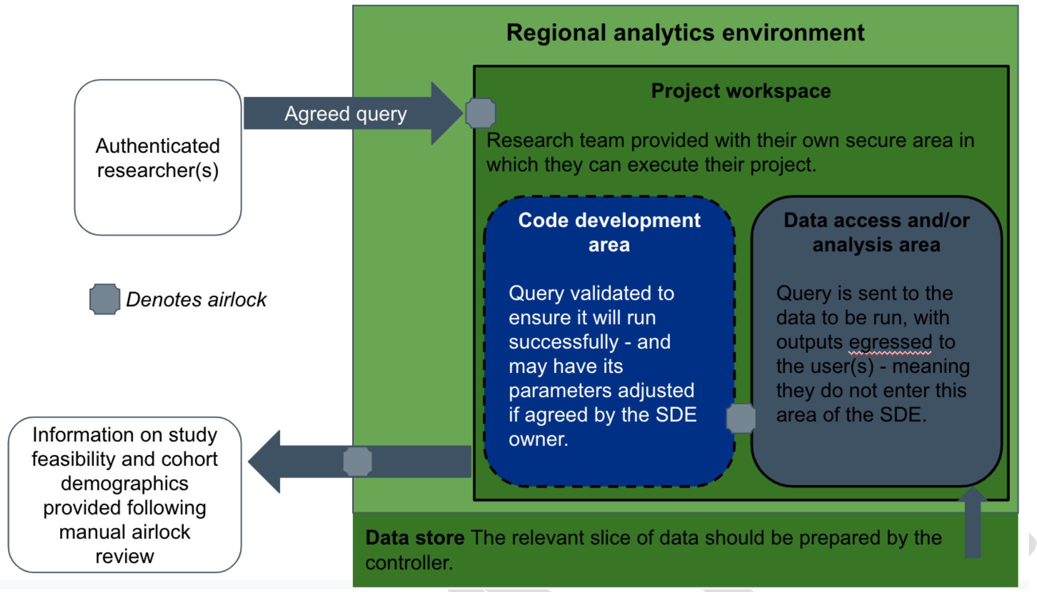
Code repository

Outputs

Outputs provided to user, or to relevant repository or dashboard.

Data store Data from multiple sources should be linked and prepared by the data controller

3. Cohort discovery



Regional analytics environment

Project workspace

Research team provided with their own secure area in which they can execute their project.

Code development area

Query validated to ensure it will run successfully - and may have its parameters adjusted if agreed by the SDE owner.

Data access and/or analysis area

Query is sent to the data to be run, with outputs egressed to the user(s) - meaning they do not enter this area of the SDE.

Authenticated researcher(s)

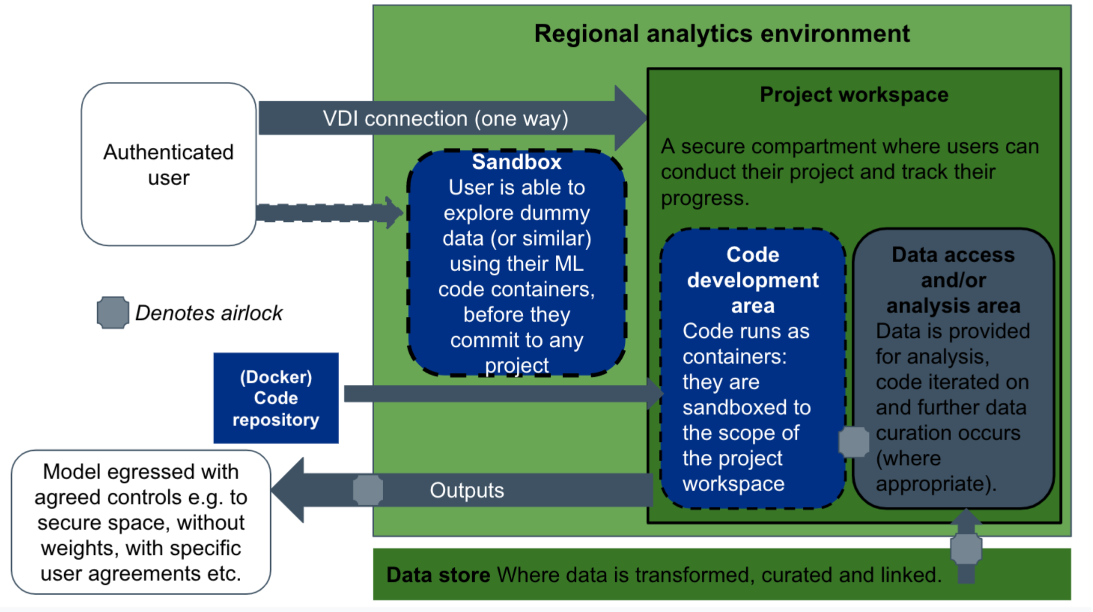
Agreed query

Denotes airlock

Information on study feasibility and cohort demographics provided following manual airlock review

Data store The relevant slice of data should be prepared by the controller.

4. Machine Learning



Regional analytics environment

Project workspace

A secure compartment where users can conduct their project and track their progress.

Authenticated user

VDI connection (one way)

Sandbox

User is able to explore dummy data (or similar) using their ML code containers, before they commit to any project

Code development area

Code runs as containers: they are sandboxed to the scope of the project workspace

Data access and/or analysis area

Data is provided for analysis, code iterated on and further data curation occurs (where appropriate).

Denotes airlock

(Docker) Code repository

Outputs

Model egressed with agreed controls e.g. to secure space, without weights, with specific user agreements etc.

Data store Where data is transformed, curated and linked.