Balancing Intellectual Property & Innovation

Who Owns Data

Dr. Jonny Pearson
Lead Data Scientist
Data Science Team
NHS England
jonathanpearson@nhs.net

Views in this slide pack are not expressive of NHS England policy.
Views are personal from my experience as an expert in healthcare innovation, synthetic data and privacy enhancing technologies.

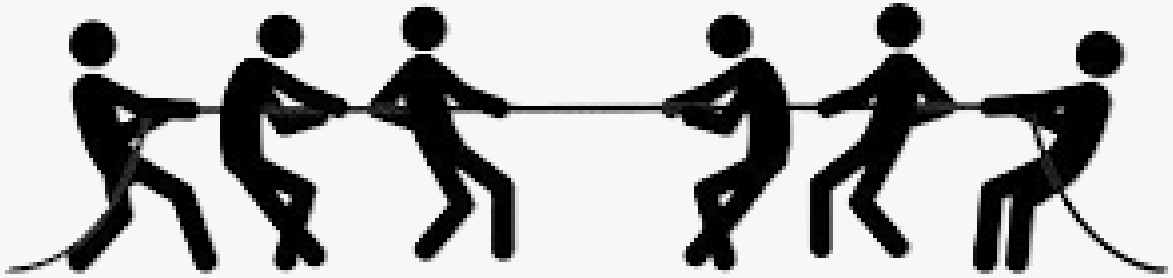
The Central Tension

Health data is simultaneously a patient right, a public asset, and a commercially sensitive resource

Irreplaceable asset
NHS holds 70+ years of longitudinal, whole-population health data – this is immensely valuable

Patient privacy must hold
Without genuine privacy protection, public opt-out collapses the data asset as care.data showed

Life-saving potential
AI-driven drug discovery, diagnostics and pathway optimisation could transform patient outcomes at scale



Commercial IP must be clear
Industry will not invest in UK data partnerships if IP ownership, licensing terms and benefit-sharing are uncertain

Innovation

Intellectual Property

Public good mandate
Research on NHS data has already produced treatments, vaccines and clinical insights that benefit millions

Confidential data must stay confidential
Trade secrets, regulatory data and clinical trial results need protection or competitive incentive to share dissolves

---

Handle privacy well
→ patients permit use

Handle benefit & governance well
→ NHS participates & public trusts

Handle IP well
→ industry invests

It’s simple

Why don’t we just ….

Get Consent

Wrap it in good enough security in the cloud

Anonymise it

Self-certify on central repository

Federate it

Open Source the model

Regulatory and Strategy Landscape

▲ WHAT IS PUSHING FOR DATA USE

MHRA Innovation Accelerator
- Fast-track routes for novel medical products & AI as medical devices
- National Commission on AI in Healthcare (2025–26)
- Driving demand for real-world evidence from NHS data

UK Life Sciences Vision
- NHS as preferred global partner for life sciences R&D
- Commits to 'data for R&D' partnerships at scale
- HDRUK established as national health data research cornerstone

Accelerated Access Collaborative
- NHS England body removing barriers to innovation adoption
- Rapid uptake pathways – links innovators to NHS data environments
- Coordinates NICE, MHRA, NHS Supply Chain & industry

Goldacre Review (2022) – Better, Broader, Safer
- Foundational TRE framework – federated access as default
- Open working, data minimisation & reproducibility as core principles
- Basis for NHS SDEs, HDRS architecture & TREvolution programme

▼ WHAT GOVERNS HOW IT IS USED

HARD LAW

UK GDPR	Lawful basis (6 options), special category health data, purpose limitation, data minimisation, subject rights incl. right to erasure. Cornerstone of all health data processing.
Data Protection Act 2018	Domestic implementation of GDPR; Schedule 3 research exemptions; Caldicott Principles enacted; sets out derogations for health & social care.
Data (Use & Access) Act 2025	Recognised Legitimate Interests as new lawful basis for health research – directly addresses consent limitations. ISMS, smart data schemes, updated research exemptions. In force June 2025.
Common Law Confidentiality	Patient data shared in clinical confidence – duty persists even after anonymisation. Applies independently of GDPR. NHS has positive legal duty to patients.

GOVERNANCE FRAMEWORKS & GUIDANCE

AI White Paper & MHRA/ICO/CMA

Safety, transparency, fairness, accountability. Partially superseded by AI Action Plan 2025.

Five Safes (TRE Principles)

Safe people, projects, data, settings, outputs. Basis for all NHS Secure Data Environments.

National Data Guardian & Caldicott Principles

10 Caldicott Principles. NDG oversees confidential data use across health & social care.

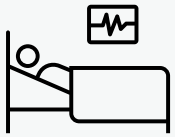
Side Note – Who owns my data?

Alex’s Story

Who has my data?

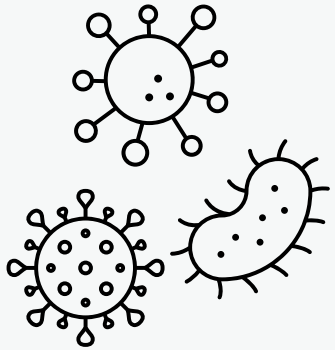
Alex was referred for ADHD and prescribed medication. Recently, Alex had an emergency admission to hospital which may have been the result of the medication. They are now on a waiting list for further management of their ADHD alongside multiple other conditions they already have including some historic mental health diagnosis. Whilst waiting for this they are frustrated by their treatment as think they have been discriminated against and don't feel well advised on how to manage their health going forwards.

Community -> Primary referral for Chronic condition with medication



ED stay with possible causal link to medication

Waiting List

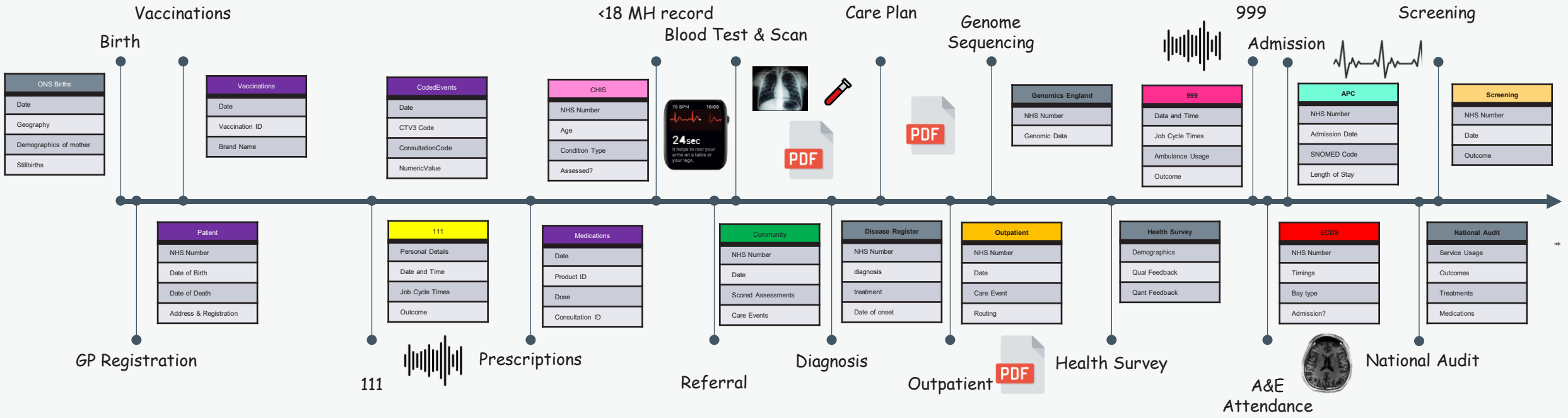


Comorbidities

Possible discrimination and dissatisfaction

Data Timeline

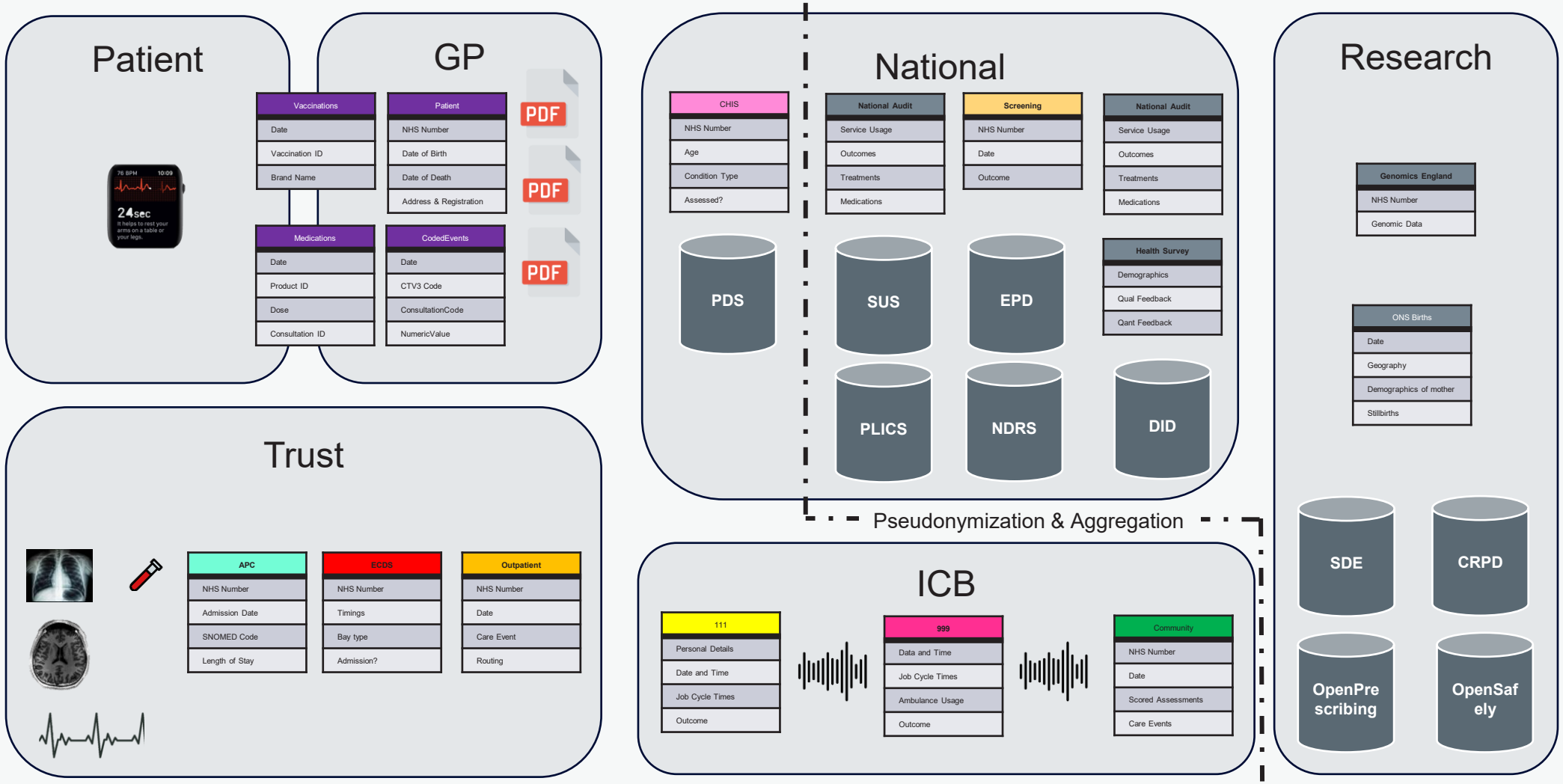
A simplistic view of data created for Alex's scenario



Pseudonymization & Aggregation

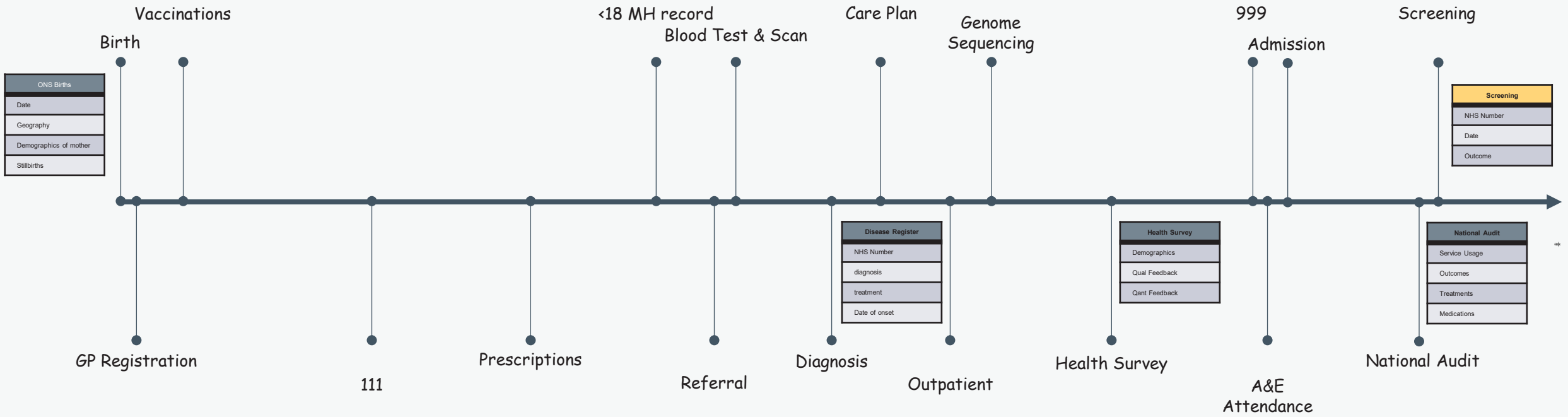
Data Timeline

A simplistic view of data created for Alex's scenario

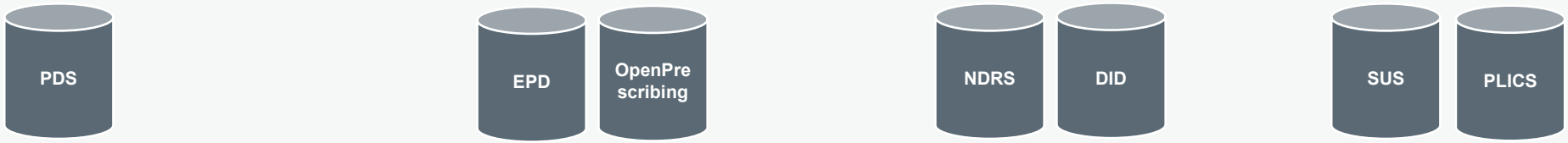


Data Timeline – National View

A simplistic view of data created for Alex's scenario

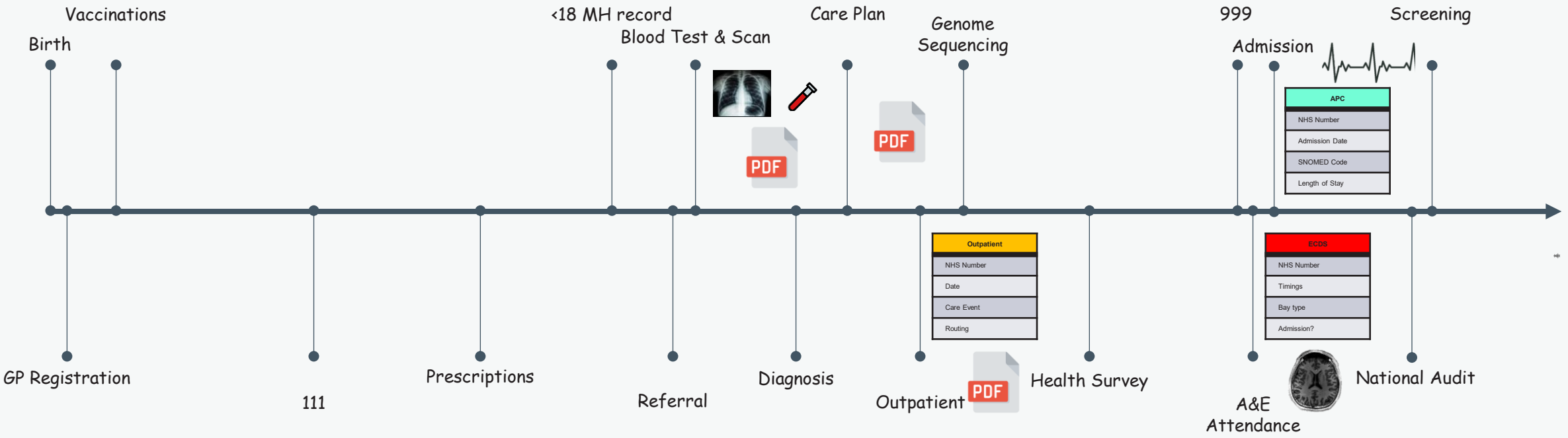


Pseudonymization & Aggregation



Data Timeline – Trust view

A simplistic view of data created for Alex's scenario



So, who is looking after my data and getting the value back for me?

What protections are there?

Proprietary Interests

It's not just one thing

The IP Ownership Chain

Raw Patient Data (not owned) → Curated Dataset (database right) → Analytical Model (copyright + patent?) → **Clinical Insight (trade secret?)** → **Marketed Product (full IP portfolio)**

The IP Family

AUTOMATIC — NO REGISTRATION

Copyright

Protects Original expression: text, software, databases, images, film

Obtained Arises on creation

Duration **Life + 70 years**

Ideas not protected — only their expression

AUTOMATIC — INVESTMENT-BASED

Database right

Protects Substantial investment in obtaining, verifying, or presenting data

Obtained Arises on creation

Duration **15 years (resets on new investment)**

Protects the data collection effort, not creativity

REGISTERED — DISCLOSURE REQUIRED

Patents

Protects Novel inventions with inventive step and industrial application

Obtained Application to IPO / EPO; formal examination

Duration **Up to 20 years**

Full public disclosure is the price of the monopoly

CONFIDENTIAL — NO REGISTRATION

Trade secrets

Protects Commercially valuable secret information (technical, business, financial)

Obtained By keeping it secret + reasonable protective steps

Duration **Indefinite (lost if disclosed)**

Reverse engineering by independent means is lawful

EQUITABLE — RELATIONSHIP-BASED

Confidential information

Protects Information with quality of confidence shared in confidential circumstances

Obtained Arises from relationship or nature of information

Duration **While confidential character is retained**

Includes patient records — core to healthcare law

REGULATORY — MEDICINES ONLY

Regulatory data protection

Protects Clinical trial data submitted for marketing authorisation

Obtained Automatic on grant of marketing authorisation

Duration **8 + 2 + 1 years**

Blocks generics from relying on originator's dossier

UK context · CDPA 1988 · Patents Act 1977 · Trade Secrets Regs 2018 · Human Medicines Regs 2012 · Copyright & Rights in Databases Regs 1997

When do these apply?

TEHDAS Data User Journey

Translated to UK context

Six Stages

01 Data discovery | 02 Permit application | 03 Data preparation | 04 Data provision | 05 Data use | 06 Results output

Seven Data Roles

Data user / researcher
The person or team seeking access to data. Responsible for completing applications, conducting approved analysis, and publishing results.

Data custodian / controller
The organisation holding and controlling the data (e.g. NHS England, SAIL, CPRD, NHS Trust). Responsible for data quality, preparation, and release.

Committee / governance
Research Ethics Committees (REC), Health Research Authority (HRA). Provide independent ethics review and grant NHS research permission.

IG / legal / DPO
Information governance leads, data protection officers, and legal teams at researcher and custodian organisations. Prepare DPIAs and DSAs.

Regulator / approver
Bodies with statutory approval powers: NHS England DARS, Confidentiality Advisory Group (CAG). Issue approvals and execute data sharing agreements.

Platform / TRE operator
Operators of Trusted Research Environments and Secure Data Environments (NHS England SDE, SAIL, CPRD, OpenSAFELY). Provision access and check outputs.

Data analyst / engineer
Technical staff who build pipelines, conduct analysis, and manage data within the secure environment. May be embedded with researcher or custodian team.

01

Data discovery

Finding and assessing available datasets before any formal application

Key Activities in UK Context

- Search HDR UK Gateway, NHS England Data Catalogue, SAIL, CPRD, UK Biobank
- Review dataset documentation, data dictionaries, and known quality issues
- Assess variable availability, temporal coverage, and population representativeness
- Identify linkage potential across datasets and data custodians
- Consult data custodians informally on feasibility before formal application
- Determine whether a Trusted Research Environment (TRE/SDE) is required

Key UK resources: HDR UK Innovation Gateway · NHS England Data Catalogue · SAIL · CPRD · UK Biobank

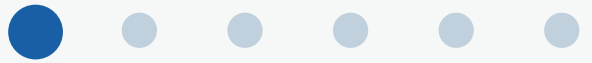
Data Roles Involved

Data user / researcher
Leads discovery; defines research question and data requirements

Data custodian
Responds to feasibility queries; maintains catalogue entries

HDR UK / Gateway
Provides federated metadata search across UK health data assets

Data scientist
Advises on technical feasibility; reviews data dictionaries



01

Data discovery

Finding and assessing available datasets before any formal application

Open Data Rights Language (ODRL)

```
<?xml version="1.0" encoding="UTF-8" ?>
<o-ex:rights xmlns:o-ex="http://odrl.net/1.1/ODRL-EX"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xmlns:o-dd="http://odrl.net/1.1/ODRL-DD"
        xsi:schemaLocation="http://odrl.net/1.1/ODRL-EX ../schemas/ODRL-EX-11.xsd
        http://odrl.net/1.1/ODRL-DD ../schemas/ODRL-DD-11.xsd">
  <o-ex:asset>
    <o-ex:context>
      <o-dd:uid>urn:ebook.world/999999/ebook/rossi-000001</o-dd:uid>
      <o-dd:name>Why Cats Sleep and We don't</o-dd:name>
    </o-ex:context>
  </o-ex:asset>
  <o-ex:permission>
    <o-dd:print>
      <o-ex:constraint>
        <o-dd:count>3</o-dd:count>
      </o-ex:constraint>
    </o-dd:print>
  </o-ex:permission>

  <o-ex:party>
    <o-ex:context>
      <o-dd:name>Alice</o-dd:name>
    </o-ex:context>
  </o-ex:party>
</o-ex:rights>
```

Alice has the rights to print ”Why cats sleep and we don’t” 3 times

https://www.w3.org/TR/odrl-model/

IP Involved

AUTOMATIC — NO REGISTRATION

Copyright

Protects Original expression: text, software, databases, images, film

Obtained Arises on creation

Duration **Life + 70 years**

Ideas not protected — only their expression

AUTOMATIC — INVESTMENT-BASED

Database right

Protects Substantial investment in obtaining, verifying, or presenting data

Obtained Arises on creation

Duration **15 years (resets on new investment)**

Protects the data collection effort, not creativity

02

Permit Application

Multiple parallel governance processes – typically the longest stage

Key Activities in UK Context

- Complete DPIA with home institution IG team and DPO
- Submit Data Access Request via NHS England DARS, SAIL, CPRD, or trust IG office
- Apply to HRA for research ethics and NHS permission where required
- Seek CAG / Section 251 support if identifiable data needed without consent
- Execute Data Sharing Agreement or Data Processing Agreement
- Obtain Caldicott Guardian sign-off at data-holding organisation

Typical UK timescales: DARS 3–6 months · HRA/REC 2–4 months · CAG 3–6 months · processes run in parallel

Data Roles Involved

Data user / researcher
Submits all application documentation; named applicant

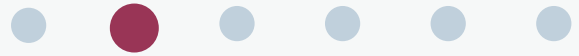
IG lead / DPO
Prepares DPIA; advises on legal basis; reviews DSA

NHS England DARS
Reviews and approves national data access requests

HRA / REC
Issues ethics opinion; grants NHS permission

CAG
Reviews Section 251 applications; advises SoS

Caldicott Guardian
Signs off patient data flows at each NHS organisation



02

Permit Application

Multiple parallel governance processes – typically the longest stage

Templated & Smart Contracts

- Pre-negotiated NHS standard data access agreements reduce legal friction
- Smart contracts (blockchain/DLT) automate compliance checking at point of access
- Auto-trigger licensing fees, royalties or benefit-sharing clauses
- Immutable audit trail — chain of custody for sensitive data assets

IP Involved

REGULATORY — MEDICINES ONLY

Regulatory data protection

Protects Clinical trial data submitted for marketing authorisation

Obtained Automatic on grant of marketing authorisation

Duration **8 + 2 + 1 years**

Blocks generics from relying on originator's dossier

CONFIDENTIAL — NO REGISTRATION

Trade secrets

Protects Commercially valuable secret information (technical, business, financial)

Obtained By keeping it secret + reasonable protective steps

Duration **Indefinite (lost if disclosed)**

Reverse engineering by independent means is lawful

AUTOMATIC — INVESTMENT-BASED

Database right

Protects Substantial investment in obtaining, verifying, or presenting data

Obtained Arises on creation

Duration **15 years (resets on new investment)**

Protects the data collection effort, not creativity

REGISTERED — DISCLOSURE REQUIRED

Patents

Protects Novel inventions with inventive step and industrial application

Obtained Application to IPO / EPO; formal examination

Duration **Up to 20 years**

Full public disclosure is the price of the monopoly

03 Data preparation

Data custodian prepares the approved extract within or adjacent to the TRE

Key Activities in UK Context

- Apply pseudonymisation / de-identification pipeline as specified in approved application
- Perform record linkage via trusted third-party using NHS number; destroy linkage key after
- Apply data minimisation — extract only variables approved in the application
- Conduct data quality checks: completeness, validity, consistency, temporal coverage
- Prepare data dictionary and provenance documentation for the extract
- Stage data in approved Secure Data Environment (SDE) workspace

UK SDEs / TREs: NHS England SDE · SAIL Databank (Wales) · CPRD (MHRA) · OpenSAFELY · Genomics England Research Environment

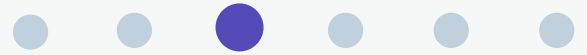
Data Roles Involved

IG lead / DPO
Oversees data flows; confirms de-identification meets approved standard

Data custodian / controller
Extracts and prepares the dataset; responsible for de-identification

TRE / SDE operator
Manages the secure environment; provisions workspace

Data Engineer
Builds and validates linkage and de-identificaiton pipelines



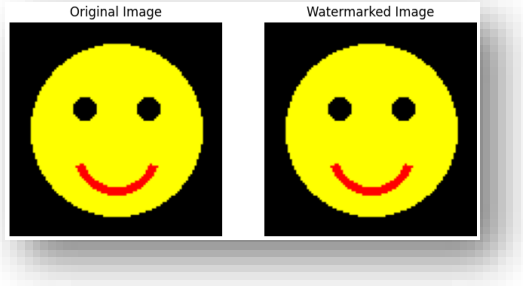
03

Data preparation

Data custodian prepares the approved extract within or adjacent to the TRE

Digital Watermarking

- Embed imperceptible markers in datasets identifying the licensed user
- Detects unauthorised redistribution or model training on stolen data
- Steganographic watermarks survive data transformation & model training
- Fingerprinting of AI model outputs traceable to training data provenance
- Limitations: fragile under heavy preprocessing; legal enforceability untested



Synthetic Data Generation

- Generating data from a statistical or data-driven model derived from a real data source or through a simulation engine.
- Supports faster access, end-to-end testing, robustness testing and bias investigations.
- Doesn't replace real data but allows a wider range of tasks to be conducted quickly
- Limitations: fidelity is dependent on the model applied and the data being modelled. Privacy thresholds are hard to demonstrate.

https://github.com/nhsengland/synthetic_clinical_notes

https://github.com/nhsengland/NHSSynth/tree/main

https://www.ihi-synthia.eu/

https://ihi-search.eu/

IP Involved

AUTOMATIC — INVESTMENT-BASED

Database right

Protects Substantial investment in obtaining, verifying, or presenting data

Obtained Arises on creation

Duration **15 years (resets on new investment)**

Protects the data collection effort, not creativity

AUTOMATIC — NO REGISTRATION

Copyright

Protects Original expression: text, software, databases, images, film

Obtained Arises on creation

Duration **Life + 70 years**

Ideas not protected — only their expression

EQUITABLE — RELATIONSHIP-BASED

Confidential information

Protects Information with quality of confidence shared in confidential circumstances

Obtained Arises from relationship or nature of information

Duration **While confidential character is retained**

Includes patient records — core to healthcare law

Stage 3 of 6

04 Data provision

Approved researcher access data within the secure environment – no copy released

Key Activities in UK Context

- Researcher accesses data only within the approved SDE — no download of patient-level data
- Access scoped to approved variables, cohort, and time period
- Researcher completes data environment induction and signs acceptable use policy
- Project workspace isolated from other projects within the TRE
- Audit logging of all data access and operations maintained by TRE operator
- Time-limited access — project end date enforced; access revoked on completion

Governed by the Five Safes framework: Safe people · Safe projects · Safe settings · Safe data · Safe outputs

Data Roles Involved

Data user / researcher
Accesses data within SDE; must comply with approved application scope

Data custodian
Retains controller responsibility; reviews access scope

TRE / SDE operator
Provisions workspace; enforces access controls and audit logging

IG lead / DPO
Monitors ongoing compliance; handles any data incidents

04

Data provision

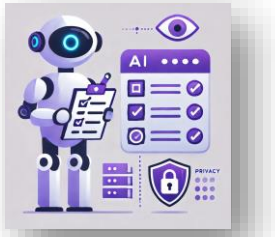
Approved researcher access data within the secure environment – no copy released

Privacy Enhancing Technologies

- Algorithmic PETs- privacy-preserving algorithms rooted in cryptography and formal privacy definitions including secure multi-party computation (SMPC), homomorphic encryption (HE), and differential privacy (DP).
- Architectural PETs- system-level methods that prevent exposure of sensitive data by design, such as federated learning (FL) and trusted execution environments (TEEs).

Intelligent Agents

- Autonomous agents monitor data use against contractual terms in real time
- Flag anomalous queries, unexpected bulk downloads, or purpose drift
- Enable dynamic consent management — update permissions without manual re-consent
- AI agents validate output against disclosure controls before release



IP Involved

AUTOMATIC — INVESTMENT-BASED

Database right

Protects Substantial investment in obtaining, verifying, or presenting data

Obtained Arises on creation

Duration **15 years (resets on new investment)**

Protects the data collection effort, not creativity

REGULATORY — MEDICINES ONLY

Regulatory data protection

Protects Clinical trial data submitted for marketing authorisation

Obtained Automatic on grant of marketing authorisation

Duration **8 + 2 + 1 years**

Blocks generics from relying on originator's dossier

05

Data use

Full analytical lifecycle conducted within the secure environment

Key Activities in UK Context

- Exploratory data analysis and data quality assessment within SDE
- Cohort construction and variable derivation as per approved analysis plan
- Statistical analysis, model development, or NLP processing within secure environment
- Iterative analysis with version control – all code retained within SDE
- Any deviations from approved analysis plan submitted as protocol amendments
- Preparation of output files for disclosure control review

UK constraint: code must be written or uploaded within the SDE; external internet access is typically blocked

Data Roles Involved

Data user / researcher
Leads analysis; responsible for adhering to approved protocol

Data custodian
Available for queries about data provenance and quality issues

TRE operator
Provides analytical tooling (R, Python, Spark) within secure environment

Data scientist / statistician
Conducts analysis; responsible for code quality and reproducibility

05

Data use

Full analytical lifecycle conducted within the secure environment

Advanced Controls & Persistent Logging

- Five Safes
- Sticky data policies - permissions travel with the data, not just the container
- Role-based query controls — limit data exposure to minimum necessary
- Immutable audit logs (blockchain anchored) of all data accesses
- Intelligent agent monitoring
- Watermarking monitoring
- PETs enablers

IP Involved

06

Results output

Outputs reviewed for disclosure risk before release

Key Activities in UK Context

- Researcher submits output files for statistical disclosure control (SDC) review
- TRE output checker reviews for small cell counts, residual disclosure, differencing attacks
- Cell suppression or aggregation applied where disclosure risk identified
- Approved outputs released to researcher outside the SDE
- Results published per open access requirements; data availability statements provided
- Code deposited in approved repository; project workspace archived or destroyed per DSA

UK output checking: minimum cell size typically n=5 or n=10 depending on environment · UKDS guidance applies

Data Roles Involved

Data user / researcher
Received approved outputs; responsible for publication

Data custodian
May review outputs for compliance with approved application scope

TRE output checker
Reviews all outputs for statistical disclosure risk before release

Ethics committee / REC
Received annual progress reports; reviews protocol amendments

06

Results output

Outputs reviewed for disclosure risk before release

Machine Unlearning

- Targeted removal of individual's data influence from a trained AI model
- Addresses UK GDPR Art. 17 Right to Erasure in AI model context
- Technical challenge: retraining costly; approximation methods emerging

https://github.com/nhsengland/priv-lm-health-extended

Disclosure Controls

- Manual or automatic checking of results.
- Statistical Disclosure Controls include suppression, aggregation, k-anonymity checking
- Privacy methods include differential privacy
- Limitations: AI model weights are an unknown egress point.

https://arxiv.org/pdf/2102.11072

IP Involved

AUTOMATIC — NO REGISTRATION

Copyright

Protects Original expression: text, software, databases, images, film

Obtained Arises on creation

Duration **Life + 70 years**

Ideas not protected — only their expression

CONFIDENTIAL — NO REGISTRATION

Trade secrets

Protects Commercially valuable secret information (technical, business, financial)

Obtained By keeping it secret + reasonable protective steps

Duration **Indefinite (lost if disclosed)**

Reverse engineering by independent means is lawful

REGISTERED — DISCLOSURE REQUIRED

Patents

Protects Novel inventions with inventive step and industrial application

Obtained Application to IPO / EPO; formal examination

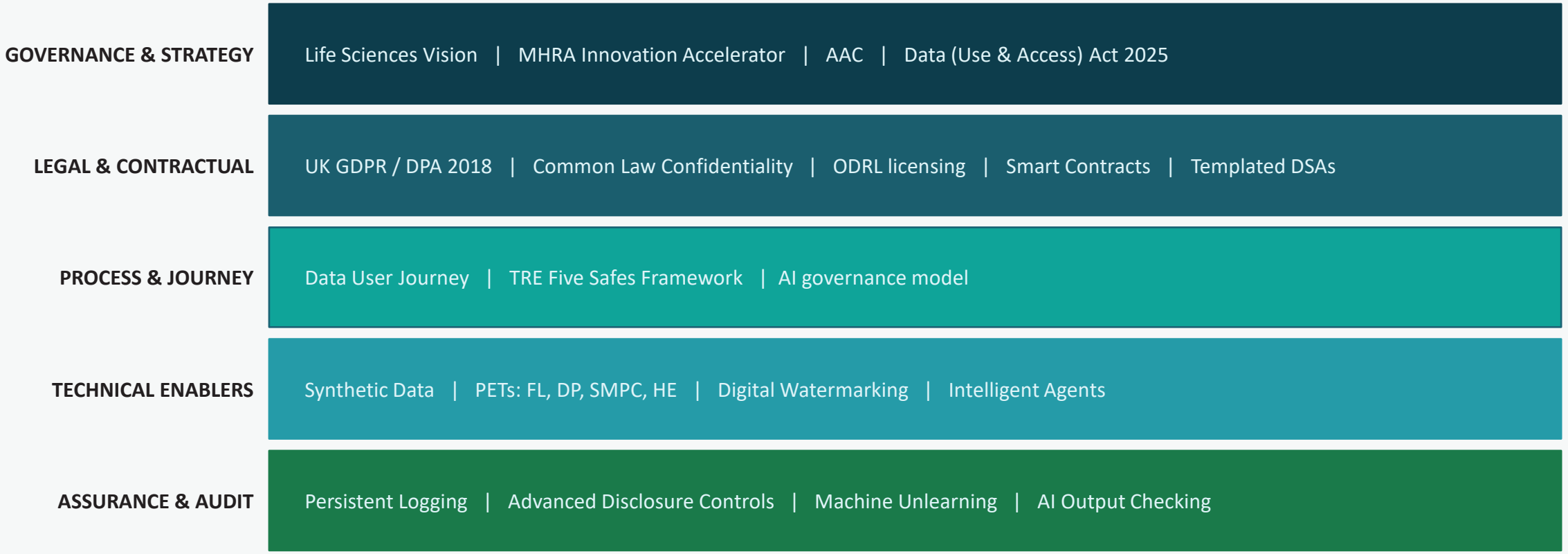
Duration **Up to 20 years**

Full public disclosure is the price of the monopoly

Bring it all together

Combination is key

A layered solution architecture for UK Health Data IP



Mandate stated -> Polic Expressed -> access controlled -> usage monitored -> outputs traced -> IP obligations enforced

Gaps and Shortcomings

Non-exhaustive

Strategic

Benefit-sharing & Value Return

No consistent mechanism to ensure NHS or patients share in commercial value created from NHS data. Model contract clauses exist but are inconsistently applied.

Patient & Public Voice in IP

Patients have no say in how downstream IP is commercialised.

Interoperability of IP Registers

No linked register of IP derived from NHS data so we cannot track what patents, trade secrets or copyrights have been created from public health data.

Process & Environment

ODRL and Smart Contracts only work with willing parties

ODRL does not enforce, only declares. Smart contracts only work in environments that adhere to these but break as soon as the data is moved to another environment.

SME Access Inequality

TRE and data access processes favour large pharma. SMEs and academics face disproportionate costs and delays. A graduated obligations framework would be fairer.

Persistent IP outside the controlled environment

A model trained in a TRE which is then exported could be used to generate synthetic data for a second model to train on. At what point does the data IP detach.

Technical

Watermarking is fragile

Steganographic watermarks can survive normal processing, but a motivated actor can bypass them

Intelligent agents only detect what they are trained on

Steganographic watermarks can survive normal processing, but a motivated actor can bypass them

Watermarking is in tension with unlearning

If machine learning erases an individual, does it also erase the associated watermark creating a conflict between privacy and traceability.

Overall: Need a legal wrapper with teeth, a provenance registry to define the chain of custody, and harmonisation

Questions to take away

Ideal - clear and shared value from data

Q1. Who should own the IP created from NHS data and what would fair look like?

Q2. Can we build technical systems sophisticated enough and fast enough to govern health data IP before the public loses trust again?

Q3. Can consent ever be truly informed given anonymisation isn't absolute and the current rate of technological development?

Q4. Should research on public health data be open by default and can we design IP frameworks that make openness the path of least resistance?