20th October, 2025

Real-World Data Pipelines in Clinical Settings: A Comprehensive Proposal for Implementing Governed, Interoperable, and Clinically Safe Data Infrastructure in the United Kingdom's Taxpayer-Funded Healthcare System

I. EXECUTIVE SUMMARY

Real-world data (RWD) generated through routine clinical care is rapidly becoming a strategic asset for state-funded healthcare systems. In the National Health Service (NHS), RWD has the potential to support service planning, quality improvement, safety surveillance, and research at a scale that is not feasible through traditional, manually assembled datasets. However, in many clinical settings, data remains fragmented across disparate systems, inconsistently defined, and difficult to access securely and reproducibly.

This whitepaper proposes a comprehensive framework for implementing real-world data pipelines in clinical settings, with the aim of producing reliable, governed, and interoperable datasets that can be safely used for operational decision-making and real-world evidence (RWE) generation. The proposed framework includes:

A reference, end-to-end pipeline architecture covering ingestion, harmonisation, terminology mapping, record linkage, quality gating, secure access, and operational monitoring;

A governance and assurance model that aligns information governance (UK GDPR and Caldicott principles), clinical safety standards for digital systems, and research governance requirements;

A pragmatic data quality scoring approach and minimum dataset specifications to ensure that RWD is fit for purpose across common clinical and research use-cases;

Three hypothetical clinical scenarios to demonstrate how robust pipelines reduce manual data collection, improve decision support, and enable auditable RWE outputs;

A phased implementation roadmap for NHS trusts and Integrated Care Systems (ICSs), including options for federated and centralised deployment models.

By adopting this framework, the NHS can reduce repeated local build effort, improve the consistency and safety of analytics, and protect patients

through clear accountability, controlled access, and transparent data provenance. Ultimately, a well-governed RWD pipeline capability supports more efficient allocation of public resources by enabling earlier insight, faster evaluation, and continuous improvement in patient care.

Y.G.

GATMEDI

18th October, 2025

II. INTRODUCTION:

Clinical services operate in an environment where demand, complexity, and resource constraints are increasing. At the same time, most modern care delivery is mediated by digital systems, generating large volumes of structured and unstructured data. The challenge is not the absence of data, but the absence of consistent, end-to-end pipelines that transform raw clinical records into trusted datasets that can be used safely for clinical operations, commissioning, and research.

In many trusts, the prevailing model relies on ad-hoc extracts, spreadsheet workflows, and project-specific data marts. These approaches are labour intensive, difficult to validate, and prone to drift when source systems, clinical pathways, or coding practices change. They also increase risk because access controls, auditability, and de-identification measures are often applied late or inconsistently.

A standardised real-world data pipeline capability is therefore a foundational investment. It enables routine production of curated datasets, reduces time-to-insight for operational teams, and supports research and innovation while maintaining public trust. This paper sets out a practical approach to building that capability in real clinical environments, where interoperability constraints, clinical safety considerations, and information governance requirements must be addressed explicitly.

III. Defining Real-World Data (RWD) and Real-World Evidence (RWE):

Real-world data refers to data relating to patient health status and/or the delivery of healthcare that is routinely collected from a variety of sources. In clinical settings, RWD typically originates from operational systems rather than controlled research environments. Real-world evidence is the clinical evidence derived from analysis of RWD, often used to understand effectiveness, safety, and outcomes in real practice populations.

Within the NHS, high-value RWD domains commonly include:

a. Electronic health record (EHR) data, including diagnoses, problems, observations, and clinical notes;

b. Administrative systems such as patient administration systems (PAS), waiting lists, and referral workflows;

c. Laboratory information systems (LIS) and results messaging, including time-stamped test results and reference ranges;

d. Radiology and imaging metadata and reports, often linked to DICOM archives (PACS) and reporting systems;

e. Medicines management systems, prescribing, dispensing, and pharmacy stock information;

f. Theatre, critical care, and device data, including high-frequency waveforms and physiological signals where available;

g. Patient-reported outcome measures (PROMs), experience measures, and pathway-specific assessments;

h. Registry, community, mental health, and social care datasets where legally and operationally feasible.

The heterogeneity of these sources is precisely why an explicit pipeline approach is required. RWD pipelines must accommodate differing event granularity, inconsistent terminology, missingness, and varying timeliness. Without rigorous standardisation and validation, analytics built on RWD can appear credible while being clinically misleading.

IV. The Case for Standardised RWD Pipelines in Clinical Settings:

In a taxpayer-funded healthcare system, a consistent approach to RWD pipelines is not merely a technical preference; it is a governance and productivity requirement. When each service, trust, or project builds bespoke integrations, the system incurs repeated cost, repeated risk, and repeated delay. Conversely, a reusable pipeline capability can generate value across multiple clinical and operational objectives.

A well-designed RWD pipeline approach can:

Protect patients by ensuring that analytics and decision support are based on validated, traceable datasets rather than ad-hoc extracts;

Improve operational efficiency by reducing manual data handling, repeated coding reconciliations, and clinician-led data collection;

Enable faster evaluation of service changes, pathway redesign, and new technologies using real-world outcomes and resource utilisation data;

Support equitable care by making it feasible to monitor variation in access, outcomes, and experience across populations and geographies;

Increase research readiness by producing curated cohorts that can be governed through secure data environments and appropriate approvals;

Strengthen accountability by providing clear audit trails, consistent definitions, and documented transformations from source to output.

It is also important to recognise that RWD pipelines, if poorly implemented, can introduce new risks. These include inappropriate access, re-identification risk, over-reliance on incomplete data, and propagation of coding artefacts into operational decisions. For this reason, pipeline design must treat information governance, clinical safety, and data quality as first-order requirements.

V. Proposed End-to-End Pipeline Architecture (Multi-Layer Model):

The proposed model is an end-to-end pipeline that separates raw ingestion from curated, governed outputs. This approach supports repeatability, controlled access, and incremental improvement without disrupting operational continuity. The key components are as follows:

1. Source Systems and Data Inventory

A formal inventory should be maintained for all source systems, including system owners, data domains, interface types, refresh frequency, and known data quality limitations. Critical dependencies (for example, laboratory result feeds driving safety surveillance) should be explicitly identified and managed as operational services rather than discretionary projects.

2. Ingestion and Integration Layer

a. Support multiple ingestion modes, including HL7 v2 messaging, FHIR APIs, batch extracts, and event streaming where available;

b. Implement change capture and incremental loading to avoid full refresh cycles and reduce strain on clinical systems;

c. Preserve source timestamps and message identifiers to support auditability and temporal analyses;

d. Treat interface mapping and transformation logic as versioned assets, subject to change control.

3. Harmonisation, Terminology, and Standardisation

a. Apply consistent terminology mapping to clinically meaningful standards (for example SNOMED CT, ICD-10, OPCS-4, dm+d, and LOINC where applicable);

b. Implement a common data model (CDM) appropriate to the intended use, noting that research-focused CDMs (such as OMOP) and operational models can coexist if governance is clear;

c. Maintain a metadata catalogue capturing field definitions, transformations, and lineage to support reproducibility and clinical interpretation;

d. Separate curated “gold” outputs from raw and intermediate layers to prevent accidental misuse of partially processed data.

4. Identity Resolution and Record Linkage

a. Where lawful and operationally appropriate, use a consistent patient identity approach (for example deterministic matching on NHS number with secondary checks);

b. Document linkage methods and confidence to prevent inappropriate use of uncertain matches in safety-critical analyses;

c. Support cross-setting linkage (acute, community, mental health) through a governed linkage service rather than one-off merges.

5. Data Quality and Validation Gates

a. Implement automated quality checks for completeness, conformance, plausibility, and timeliness at each pipeline stage;

b. Define clinical plausibility rules jointly with clinical leads (for example impossible laboratory values or conflicting demographics);

c. Establish exception handling workflows so that data issues are triaged, corrected where appropriate, and communicated to downstream users;

d. Maintain a “known issues” register so that analysts and clinicians understand limitations and avoid incorrect inference.

6. Secure Storage, Compute, and Controlled Access

a. Use tiered environments (raw, curated, and approved outputs) with role-based access control and least-privilege principles;

b. Implement encryption in transit and at rest, comprehensive audit logging, and routine access review;

c. Provide a secure analytics environment (Trusted Research Environment or equivalent) for approved research and evaluation activities;

d. Ensure that data exports are governed, minimised, and traceable, with clear separation between identifiable, pseudonymised, and anonymised outputs.

7. Analytics Delivery and Operational Use

a. Provide standard outputs for operational reporting (for example performance metrics, pathway dashboards, safety surveillance, and capacity planning);

b. Enable reproducible research outputs through version-controlled code, documented cohorts, and repeatable transformations;

c. Where advanced analytics or machine learning is deployed, implement model governance, monitoring for drift, and clear clinical responsibility for interpretation and action.

8. Pipeline Operations, Monitoring, and Change Control

a. Define service-level expectations (refresh frequency, latency, uptime) for high-value datasets and feeds;

b. Monitor pipelines for failures, schema drift, data drift, and unusual patterns that may reflect upstream process change;

c. Apply structured change control to interface updates, mapping changes, and release cycles to reduce clinical disruption;

d. Establish incident response processes aligned with both information security and clinical safety governance.

VI. Information Governance, Privacy, and Public Trust:

RWD pipelines must be designed within a clear information governance (IG) framework. In the UK context, this includes compliance with data protection law, alignment with Caldicott principles, and adherence to NHS security standards. In practical terms, IG should be operationalised as a set of design constraints and decision points rather than treated as a final approval step.

Key IG requirements for clinical RWD pipelines include:

a. Defined purpose and lawful basis, with explicit documentation of whether use is direct care, operational management, service evaluation, or research;

b. Data minimisation by default, ensuring that datasets contain only the fields required for the approved purpose;

c. Pseudonymisation or anonymisation where appropriate, with separation of identity data and robust key management;

d. Transparent access control, audit logging, and routine review of user entitlements;

e. Clear data sharing agreements and processor contracts where data is processed across organisational boundaries;

f. Explicit management of patient choice and transparency obligations, including appropriate handling of national opt-out where applicable;

g. Secure data environments for secondary use, ensuring that outputs are reviewed and disclosed in line with policy.

Public trust is a prerequisite for sustainable RWD use. A pipeline capability should therefore include clear documentation of safeguards, accessible explanations of how data is used, and demonstrable benefits to patient care. Trust is strengthened when governance is visible, consistent, and accountable.

VII. Clinical Safety and Assurance in Data-Driven Workflows:

In clinical environments, analytics outputs increasingly influence operational decisions and, in some cases, clinical decision-making. Where pipeline outputs are used for prioritisation, triage, alerts, or decision support, they must be managed with explicit clinical safety assurance. This is

particularly important when introducing automated rules or machine learning outputs that may not be intuitive to end users.

A pragmatic safety approach should include:

a. Clear definition of intended use, users, and decisions influenced by the dataset or analytic output;

b. Named clinical ownership for the pathway, dashboard, alert, or model that consumes pipeline outputs;

c. Hazard identification for foreseeable failure modes (for example missing data, delayed feeds, misclassification, or incorrect mapping);

d. Verification and validation processes, including user acceptance testing with clinical scenarios and edge cases;

e. Ongoing monitoring, including alert performance, false positives/negatives, and dataset drift;

f. A change management and rollback plan so that unsafe behaviour can be rapidly mitigated if detected.

This safety discipline should not be viewed as a barrier to innovation. Rather, it enables scalable deployment of analytics across multiple teams because it provides a repeatable assurance pathway and a shared language for managing risk.

VIII. Data Quality, Standard Definitions, and Reproducibility:

Data quality is a practical determinant of whether RWD pipelines deliver value or create confusion. Clinical users require confidence that measures are stable, definitions are consistent, and that changes in data reflect real clinical change rather than artefacts of system updates or coding practice. A pipeline programme should therefore adopt explicit quality metrics and publish them as part of routine reporting.

We propose a simple Data Quality Score (DQS) that can be calculated per dataset and per refresh cycle. The purpose is not to imply absolute truth, but to provide a consistent indicator of fitness for purpose and to trigger investigation when quality degrades. One pragmatic formulation is:

Data Quality Score (DQS) = min(100, (0.25 * C) + (0.25 * K) + (0.20 * P) + (0.20 * T) + (0.10 * L))

Where: C = Completeness (% required fields populated), K = Conformance (% records meeting schema and coding rules), P = Plausibility (% records passing clinical plausibility checks), T = Timeliness (% data within agreed latency), and L = Linkage Quality (% records with acceptable linkage confidence). Weights should be adjusted depending on the use-case; for

example, safety surveillance may weight timeliness more heavily than longitudinal research.

In addition to numeric scoring, reproducibility requires disciplined definition management. This includes maintaining versioned code lists, phenotype definitions, and measure specifications; documenting any changes to source systems; and ensuring that downstream dashboards and studies can be re-run against specific dataset versions.

IX. Three Hypothetical Clinical Scenarios:

The following examples illustrate how an end-to-end RWD pipeline approach can be applied in real clinical settings. These scenarios are hypothetical but representative of common operational and clinical requirements.

Scenario 1: Near Real-Time Deterioration Surveillance (e.g., Sepsis Pathway)

Objective: Combine observations, laboratory results, and clinical events to monitor time-to-treatment and identify delays in pathway execution.

Pipeline requirements: Ingestion of vital signs and lab results with timestamp preservation; terminology mapping for observations; timeliness monitoring; and a safety case for any alerts or escalation triggers derived from the dataset.

Operational output: A dashboard for pathway performance and a controlled alerting workflow where clinicians retain decision-making responsibility. Data quality focus is on timeliness (T) and plausibility (P).

Scenario 2: Medicines Safety and Adverse Event Detection (e.g., Acute Kidney Injury Risk)

Objective: Detect potential medication-related harm by linking prescribing/administration records with laboratory trends and comorbidity context.

Pipeline requirements: Medicines data harmonised to consistent identifiers; lab result standardisation; robust linkage; and explicit governance for secondary use. Rules and thresholds should be clinically validated, and false positives monitored to avoid alert fatigue.

Operational output: A safety surveillance report and, where appropriate, an MDT review workflow. Data quality focus is on conformance (K) to medication coding and linkage quality (L).

Scenario 3: Real-World Outcomes Evaluation (e.g., Elective Surgery Pathway Redesign)

Objective: Evaluate whether a pathway change improves outcomes and reduces length of stay, using routine data rather than bespoke audit collection.

Pipeline requirements: Cohort definition using standard procedure coding; linkage to outcomes and readmissions; consistent date logic; and reproducible measure definitions so that results remain comparable across time.

Operational output: A service evaluation dataset and repeatable metrics pack that can be used across trusts within an ICS. Data quality focus is on completeness (C) of key fields and stability of definitions over time.

X. Implementation Roadmap for Trusts and Integrated Care Systems:

Implementing RWD pipelines is a programme, not a single deployment. The recommended approach is to start with high-value, tractable use-cases and progressively build reusable capability. Key phases to consider include:

Phase 1: Establish the Foundations

a. Create an enterprise data inventory and prioritise datasets aligned to clinical and operational value;

b. Define governance roles, including IG leadership, clinical safety leadership, and operational service ownership;

c. Agree minimum standards for terminology, metadata, and data quality reporting.

Phase 2: Deliver a Minimum Viable Pipeline

a. Implement ingestion and curated datasets for one or two priority pathways;

b. Stand up secure access patterns and a controlled analytics environment;

c. Publish initial quality metrics and implement a defect triage process.

Phase 3: Scale and Standardise

a. Expand to additional domains (medicines, imaging metadata, outcomes, PROMs) and increase automation of quality checks;

b. Introduce a consistent common data model strategy and shared phenotype definitions where appropriate;

c. Implement monitoring for drift and establish structured change control for interfaces and mappings.

Phase 4: Federate Across an ICS (Where Appropriate)

a. Adopt a federated model where data remains within trust-controlled environments but outputs are standardised for cross-system analysis;

b. Establish shared governance, shared definitions, and shared tooling to prevent divergence;

c. Ensure that cross-organisational data access is supported by formal agreements and transparent decision-making.

A deliberate focus on repeatability and shared standards reduces long-term cost. It also supports rapid response to emerging needs, such as safety signals, winter pressures, or evaluation of new interventions, because the foundational pipelines are already in place.

XI. Conclusion:

Real-world data pipelines are now a core enabling capability for modern clinical operations and evidence generation. In the NHS, the opportunity is significant: routine data can support safer care, more efficient service delivery, and faster evaluation of change. However, without disciplined architecture, governance, and quality assurance, RWD can just as easily create noise and risk as it can create insight.

The framework proposed in this whitepaper provides a practical blueprint for building end-to-end RWD pipelines that are governed, clinically safe, and interoperable. It emphasises reusable capability, clear accountability, and published data quality measures. By investing in this approach, trusts and ICSs can reduce repeated local effort, improve analytic reliability, and ensure that secondary use of data remains aligned to public benefit and public trust.

In conclusion, a standardised pipeline capability should be treated as essential infrastructure for a taxpayer-funded system operating in a capital- and information-intensive model. Collaboration among clinical leaders, informatics teams, governance bodies, and operational stakeholders is required to translate this capability into sustained improvements in patient care and system sustainability.

REFERENCES

National Data Guardian. (2013, updated). The Caldicott Principles. Retrieved from https://www.gov.uk/government/publications/the-caldicott-principles

UK Government. (2018). Data Protection Act 2018. Retrieved from https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

Information Commissioner's Office (ICO). Guide to the UK GDPR. Retrieved from https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/

NHS Digital / NHS England. Data Security and Protection Toolkit (DSPT). Retrieved from https://www.dsptoolkit.nhs.uk/

HL7 International. (n.d.). FHIR Specification. Retrieved from https://hl7.org/fhir/

Health Level Seven International. (n.d.). HL7 Version 2 Product Suite. Retrieved from https://www.hl7.org/implement/standards/product_brief.cfm?product_id=185

Digital Imaging and Communications in Medicine (DICOM). (n.d.). DICOM Standard. Retrieved from https://www.dicomstandard.org/

OHDSI. (n.d.). OMOP Common Data Model. Retrieved from https://www.ohdsi.org/data-standardization/the-common-data-model/

World Health Organization. (2019). ICD-10: International Statistical Classification of Diseases and Related Health Problems. Retrieved from https://icd.who.int/browse10/2019/en

NHS England. (2022). Data Saves Lives: Reshaping health and social care with data. Retrieved from https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data