Clinical Data Management for Myotonic Dystrophy Type 1

A Three-Month Research Action Plan - IGTP

Gary Espitia - espitiagmd@gmail.com

February 2026

This document presents a three-month action plan to design, build, and validate a minimum viable platform for prospective data collection in Myotonic Dystrophy Type 1 (DM1). The platform is conceived as mobile-first, respects a clear role-based governance model, and is aligned with the General Data Protection Regulation (GDPR) and the European Health Data Space (EHDS). For clinical interoperability it adopts Fast Healthcare Interoperability Resources (FHIR), while the research analytics layer relies on the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM). Every component in the stack is Free and Open-Source Software (FOSS) and follows a local-first philosophy, so that no proprietary dependency constrains future evolution.

Throughout the plan, key artefacts-access-control matrices, responsibility charts, timelines, risk assessments, and architectural diagrams-are referenced as clickable links that point to the full-resolution versions hosted in the **project repository**.

1 Data Quality and Uniformity Evaluation

1.1 Scope and Approach

Clinical studies in rare diseases like DM1 typically assemble multi-modal data-clinical assessments, genomic sequencing, and proteomic profiling-from several hospitals, each with its own conventions. Before any analysis can be trusted, these heterogeneous sources need to be reconciled. The quality plan therefore pursues five practical goals: detecting structural and semantic inconsistencies along with duplicates and normalisation gaps; remediating entry errors and implausible values; characterising missingness and cross-form discordances; quantifying site- and user-level variation; and keeping every cleaning step auditable and reproducible.

The process begins with a thorough data inventory that documents every table's domain, purpose, primary and foreign keys, controlled vocabularies-International Classification of Diseases 10th Revision (ICD-10), Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), Anatomical Therapeutic Chemical Classification System (ATC), Logical Observation Identifiers Names and Codes (LOINC)-and expected cardinalities. A codebook then aligns these documented expectations with the reality found during profiling.

The working environment centres on **PostgreSQL** as the relational hub, chosen because it enforces referential integrity, schema types, and Atomicity–Consistency–Isolation–Durability (ACID) transactions at the engine level-properties that are critical when linking clinical, genomic, and proteomic tables and that align naturally with both the FHIR facade and the OMOP CDM. Automated validation is implemented in **Python** (pandas, pandera, pytest), operational dashboards in **Grafana**, and domain-specific processing with tools such as bcftools (variant calling), plink (genotype quality control), and MSstats (proteomic quantification). The entire analytical toolchain-R, Python, pandoc, and LaTeX-is version-pinned through a **Nix** flake, so that any collaborator can reproduce the exact working environment from a single declarative configuration. On the OMOP instance, **Observational Health Data Sciences and Informatics (OHDSI)** tools (Achilles, DataQualityDashboard) provide an additional, standardised quality lens.

1.2 Quality Dimensions and Checks

Structural integrity is the first line of defence: primary key uniqueness, foreign key validation, type and format parsing, and code validation against SNOMED CT, ICD-10, ATC, and LOINC reference tables. All of these checks are expressed as pandera schemas and pytest suites kept under version control.

Uniformity covers unit harmonisation (with documented conversion factors), category mapping to canonical representations, and identifier format normalisation. Every transformation retains the original value and is logged.

Duplicate detection follows a four-tier strategy. First, exact hash matching on composite keys (subject identifier, visit date, form type) catches verbatim duplicates. Second, a fuzzy blocking step groups records that share phonetic or edit-distance similarity on name, date of birth, and site, reducing the comparison space. Third, probabilistic scoring with the recordlinkage or Dedupe libraries assigns a match probability to each candidate pair; pairs above a configurable threshold are merged automatically, while borderline pairs are routed to a manual adjudication queue with side-by-side display. Fourth, cross-modal concordance checks-genotype concordance via plink and proteomic intensity correlations-detect cases where the same biological sample was registered under different clinical identifiers.

Missingness is characterised at the variable level, stratified by site, personnel, and time. Heatmaps make clustered patterns visible, and the plan explicitly distinguishes design-based skip logic from true protocol deviations. Threshold-based alerts trigger formal queries.

Plausibility relies on clinically defined ranges encoded as pandera schemas, logical consistency rules (for example, age at diagnosis must not exceed age at last visit, and future dates are rejected), Interquartile Range / Median Absolute Deviation (IQR/MAD)-based outlier detection per site, and checks for data-entry artefacts such as digit heaping or unrealistic entry speeds.

Cross-form and cross-modal discordances are caught through automated reconciliation rules executed after each data import. Baseline diagnoses recorded on the eligibility form are compared against the longitudinal problem list; medications listed on the concomitant-medication log are cross-referenced with prescriptions captured in the e-pharmacy module; temporal consistency of key dates (symptom onset, sample collection, consent withdrawal) is enforced by constraint checks that flag any implausible ordering. As a final biological cross-check, clinically recorded sex is compared against genetically inferred sex from genomic data, and any mismatch triggers a priority query for site review.

Finally, **site and personnel performance** is tracked through a set of Key Performance Indicators (KPIs): field-level completeness rates, edit-check failure rates, median query resolution time, and error proxies such as correction frequency per operator. These KPIs are exposed in Grafana dashboards, shared with the team in weekly status reports, and used to prioritise retraining or process adjustments so that adverse patterns can be addressed early.

1.3 Correction, Cleaning, and Traceability

No value is ever overwritten silently. Every correction records original and new values, the actor, a timestamp, and a reason in an append-only audit table. A centralised query log-carrying a unique identifier, patient/sample/site reference, severity, assignee, and status transitions-structures the resolution workflow. Standard Operating Procedures (SOPs) distinguish routine corrections from high-impact changes that require Principal Investigator (PI) approval and from sample-level Quality Control (QC) decisions. Cleaned datasets are versioned with data-cut dates and git tags. After FHIR and OMOP mapping, quality is reassessed using dedicated validators and the OHDSI DataQualityDashboard.

2 Proposal for Dynamic Analysis

2.1 Architecture

To ensure that research queries never slow down clinical data capture, the architecture maintains two separate PostgreSQL instances: one operational (data entry) and one analytical (OMOP CDM). Keycloak, an open-source identity provider, issues tokens that carry role and site claims, so every query runs under governance-aware access control and is logged. All queries, cohort definitions, and report templates live in version control.

2.2 Fast Queries

Performance is achieved through composite B-tree indexes on the OMOP tables (combining `person_id` with date fields, and indexing `concept_id`), as well as on genomic (sample, gene) and proteomic (protein accession) columns. Three materialised views, refreshed nightly, serve the most common access patterns: a patient-level summary that joins all modalities in a single query, a longitudinal measurements view for time-series exploration, and a site-level completeness view for operational monitoring. PgBouncer handles connection pooling, and expensive analytical queries are routed to a streaming read replica.

2.3 Exploration and Correlation

Because the patient entity serves as the link across all modalities, a library of parameterised SQL and Python templates has been prepared for the most common cross-modal queries. Researchers work in containerised JupyterLab or RStudio Server environments that connect through read-only, role-scoped credentials and come pre-loaded with analytical libraries (pandas, scikit-learn, lifelines, tidyverse, survival, DESeq2, MSstats). Environment definitions are managed as Nix flakes, so that every dependency version is locked and any analysis session can be reproduced exactly. For investigators who prefer a graphical interface, OHDSI Atlas offers point-and-click cohort definition with JSON export for downstream scripts. Modular workflows cover survival analysis, mixed-effects models, clustering, and predictive classifiers, all with logged metadata for reproducibility.

2.4 Reports and Dashboards

Grafana serves operational needs: enrolment progress, completion rates, missingness heatmaps, query volumes, and site KPI scorecards, all defined as code and scoped by role so that each team member sees the indicators most relevant to their function. **Apache Superset** supports research-oriented exploration with interactive charts, ad-hoc filtering, and row-level security governed by OAuth 2.0 tokens issued through Keycloak. Scheduled reports are orchestrated by Airflow or Prefect Directed Acyclic Graphs (DAGs), which render parameterised scripts to PDF or HTML via Quarto, each tied to a specific script version and data cut.

The platform is designed for adaptability: schema evolution is handled through Alembic or Flyway, Extract–Transform–Load (ETL) pipelines are configuration-driven (JSON/YAML), and analytical components are modular and independently replaceable. New modalities-wearable sensors, imaging-and external sources such as disease registries or molecular databases can be integrated without rearchitecting.

2.5 Exploratory Statistical Study

To illustrate the analytical capabilities in practice, an exploratory study was carried out using synthetic data from the Clinical Data Interchange Standards Consortium (CDISC). Starting from demographic summaries, it progresses through longitudinal vital-sign trajectories and change-from-baseline laboratory analyses to Kaplan–Meier survival curves. The interactive report,

including all code and figures, is available at **Exploratory Analysis - GitHub Pages**.

3 Mobile App for Prospective Data Collection

3.1 Architecture and Technologies

The system is organised in seven clearly separated layers: (1) a patient-facing mobile app together with a clinician portal; (2) a RESTful API layer built on FastAPI (Python) or NestJS (TypeScript), which acts as the single gateway between every client application and the back-end-validating incoming payloads against FHIR profiles, enforcing authorization tokens issued by Keycloak, orchestrating writes to the database, and exposing versioned endpoints for the mobile app, the clinician portal, and third-party integrations alike; (3) Identity and Access Management (IAM) through Keycloak, providing OAuth 2.0 authorization flows and OpenID Connect for federated identity, Role-Based Access Control (RBAC), and Multi-Factor Authentication (MFA); (4) an operational PostgreSQL database that stores only pseudonymised data; (5) de-identification pipelines orchestrated by Airflow or Prefect; (6) the OMOP CDM research warehouse with OHDSI Atlas; and (7) an observability stack comprising Prometheus, Grafana, Loki, and an append-only audit log. Everything is containerised with Docker or Podman, vendor-independent, and fully open-source. A zoned view of all components and their interactions is provided in the **Architecture Map**, and the data-flow perspective in the **Data Flow Diagram**.

The mobile app, built with Flutter or React Native, is designed to work offline-first: data is stored in an encrypted SQLite database on the device and synchronised over HTTPS in the background when connectivity returns. Given the fatigue that DM1 patients commonly experience, the interface emphasises save-and-return capability, large tappable controls, and accessibility features. The clinician portal (React + TypeScript) provides site-scoped data review, annotation tools, query management, and adherence dashboards. A dedicated data manager console allows configuration of forms, edit checks, and data locks.

3.2 Data Integrity

Every record entering the system is validated against FHIR Release 4 (R4) profiles. Twelve DM1-specific profiles have been defined-Patient, RelatedPerson, Encounter, Questionnaire, QuestionnaireResponse, Observation, Condition, MedicationStatement, MedicationRequest, AdverseEvent, Consent, and AuditEvent-ensuring that only well-formed, terminologically coded data is accepted. Edit checks run at both the client and the server. Consent is verified at every ingestion event, and an append-only audit trail records every action. Data can be locked at the subject, visit, or study level. Pseudonymisation is embedded by design: direct identifiers reside only in a network-segmented Identity Store. Standard terminologies-SNOMED CT, LOINC, ATC, ICD-10, Orphanet (ORPHA:273)-are complemented by OMOP vocabularies loaded through Athena. Concept mapping is facilitated by the OHDSI suite: WhiteRabbit for source profiling, Rabbit in a Hat for mapping specification, and Usagi for vocabulary alignment.

3.3 Iteration with the Clinical Team

Designing for a rare-disease population demands close collaboration with the people who will actually use the system. In Month 1 the team works with investigators and patients on paper prototypes. Month 2 introduces structured walkthroughs in a test environment. Month 3 brings usability sessions with DM1 patients and site staff, whose feedback is actioned before the pilot. Throughout the project, weekly Clinical Data Management (CDM)–development meetings, bi-weekly data engineering alignment sessions, and monthly governance reviews keep all parties synchronised.

4 Regulatory Compliance and Security

4.1 GDPR and EHDS

The platform's lawful basis for processing health data rests on explicit electronic consent (eConsent) under Articles 6(1)(a) and 9(2)(a), complemented by public health and research bases (Articles 9(2)(i) and 9(2)(j)). Data minimisation is enforced at the protocol level: only what the study requires is collected, and optional data elements demand separate consent. Subject rights-access, rectification, erasure, and withdrawal-are supported through documented, tested procedures. A Data Protection Impact Assessment (DPIA) is structured under Article 35. EHDS alignment is achieved through FHIR + OMOP interoperability, a clear separation of operational and research environments, transparent logging of any secondary use, and cross-border data-transfer safeguards.

4.2 Governance, Traceability, and Access Control

Every action in the system is recorded in an append-only audit trail that captures the actor, role, action type, the object before and after modification, a UTC timestamp, device and IP information, and a mandatory free-text reason for corrections. Consent is granular (core study, optional elements, secondary research, external sharing) with versioned forms and digital signatures. When a participant withdraws, data collection stops immediately and downstream pipelines are flagged.

Access is governed through RBAC enforced at two levels-Keycloak (which issues OAuth 2.0 tokens carrying role and site claims) and the application itself-covering 7 roles and 38 permissions across 11 categories. The full permission breakdown is shown in the **RBAC Permission Matrix**. The guiding principles are least privilege, site-level scoping, quarterly access reviews, and time-limited elevated permissions that require explicit approval.

4.3 Encryption and Incident Response

All data in transit is protected by Transport Layer Security (TLS) 1.2 or higher, with mutual TLS (mTLS) for internal service communication. Data at rest is encrypted at the PostgreSQL volume level, supplemented by pgcrypto for sensitive fields and encrypted SQLite on mobile devices. A documented incident response procedure covers containment, impact assessment, GDPR-mandated notification within 72 hours (Articles 33–34), remediation, and post-incident review.

4.4 Risk Assessment

Four key project risks have been identified and assessed. Their likelihood (L), impact (I), and planned mitigations are summarised below. Likelihood is rated as **H**igh or **M**edium; impact is rated as **Maj**or or **Mod**erate. The visual matrix is available at **Risk Likelihood–Impact Matrix** and the detailed narrative at **Risk Assessment**.

Risk	L / I	Mitigation
R1: FHIR/OMOP mapping complexity	H / Maj	Prioritise a minimal variable subset first; phase broader coverage
R2: Suboptimal DM1 usability	M / Maj	Early and repeated usability testing; accessible, fatigue-aware design
R3: Slow governance/legal reviews	H / Mod	Initiate in Month 1; assign named reviewers with agreed turnarounds
R4: FOSS integration complexity	M / Mod	Stand up an integration environment from week 1; test incrementally

5 Work Organisation, Prioritisation, and Timeline

The work is structured in four overlapping workstreams distributed across three months, as shown in the **Project Gantt Chart**.

Month 1 - Foundation and Design. The first month focuses on establishing the study data specification-capturing every data element, instrument, and schedule-and placing it under version control. The coding strategy is defined (SNOMED CT, LOINC, and OMOP mappings), together with a minimal set of FHIR profiles. A draft RBAC matrix is prepared, and initial governance documents are started: a Data Management Plan (DMP) outline, the DPIA structure, consent requirements, and data flow diagrams.

Month 2 - Configuration and Implementation. Questionnaires, schedules, and edit checks are configured in the test environment. Mobile and portal prototypes go through structured walkthroughs with the clinical team. FHIR mappings are validated against sample FHIR bundles. The OMOP ETL design is initiated using WhiteRabbit and Rabbit in a Hat. Audit trail and correction workflows are implemented. Governance documents-DPIA, retention rules, consent language-are refined.

Month 3 - Validation and Pilot. Formal testing covers typical and edge-case DM1 scenarios. RBAC and audit trail behaviour is verified. End-to-end FHIR-to-OMOP validation is performed through Atlas cohort definitions. Usability sessions with DM1 patients and site staff take place, and findings are actioned. The DMP and governance documentation are finalised. The pilot scope is defined: participating sites, enrolment targets, monitoring procedures, and escalation paths.

Key milestones: data specification complete (end of March); test environment fully configured (end of April); OMOP ETL validated (mid-May); pilot-ready (end of May).

The responsibility distribution follows the **RACI Responsibility Matrix**, which maps eight roles across all deliverables: the CDM role is Responsible and Accountable for most study design, configuration, testing, and documentation tasks. Clinical Investigators are Consulted on data specification, instrument design, and usability. Development and Technical leads drive FHIR profile creation and app prototyping. Data Engineering is responsible for the OMOP ETL pipeline. Biostatistics is Consulted on data specification and analytical workflows. The Legal / Data Protection Officer (DPO) function is accountable for the DPIA. Study Leadership steers pilot preparation. DevOps / SysAdmin supports RBAC verification, audit trail infrastructure, and deployment.

Day-to-day coordination relies on weekly CDM–Development syncs, bi-weekly CDM–Data Engineering alignment meetings, and monthly governance reviews. The DM1-specific interface design-simplified screens, a proxy mode for caregivers, save-and-return, and configurable reminders-is refined continuously based on feedback. Interoperability targets extend beyond the immediate project to include European Reference Network for Neuromuscular Diseases (EURO-NMD) registries (Orphanet, Human Phenotype Ontology [HPO]), EHDS secondary-use pathways, and CDISC as a secondary mapping target for regulatory submissions.

Full repository: **https://github.com/stradichenko/task-ref-2026-14**