Clinical Data Management for Myotonic Dystrophy Type 1

A Three-Month Research Action Plan — IGTP

Gary Espitia

February 2026

This plan outlines the path from concept to a validated minimum viable platform for prospective data collection in Myotonic Dystrophy Type 1 (DM1). The platform is mobile-first, role-based, aligned with the General Data Protection Regulation (GDPR) and the European Health Data Space (EHDS), and uses Fast Healthcare Interoperability Resources (FHIR) for clinical interoperability alongside the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) for research analytics. The entire stack is Free and Open-Source Software (FOSS) and local-first. Supporting artefacts—Role-Based Access Control (RBAC) matrix, Responsible–Accountable–Consulted–Informed (RACI) matrix, Gantt chart, risk matrix, data flow diagram—are linked throughout and collected in the **companion annex**. The full repository is available at **github.com/stradichenko/task-ref-2026-14**.

1 Data Quality and Uniformity Evaluation

1.1 Scope and Approach

The quality plan targets a multi-modal DM1 dataset (clinical, genomic, proteomic) collected across multiple hospitals. The aims are: (1) detect structural and semantic inconsistencies, duplicates, and normalisation gaps; (2) remediate entry errors and implausible values; (3) characterise missingness and cross-form discordances; (4) quantify site- and user-level variation; (5) keep all cleaning auditable and reproducible.

Preparation starts with a data inventory documenting every table's domain, purpose, keys, 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 aligns documented expectations with profiled reality. The working environment uses **PostgreSQL** as the relational hub, **Python** (pandas, pandera, pytest) for automated validation, **Grafana** for site-level dashboards, domain tools (bcftools, plink, MSstats), and **Observational Health Data Sciences and Informatics (OHDSI)** tools (Achilles, DataQualityDashboard) on the OMOP instance.

PostgreSQL is preferred because it enforces referential integrity, schema types, and Atomicity–Consistency–Isolation–Durability (ACID) transactions at the engine level—critical when linking clinical, genomic, and proteomic tables—and aligns directly with both the FHIR facade and OMOP CDM.

1.2 Quality Dimensions and Checks

Structural integrity. Primary key uniqueness; foreign key validation; type/format parsing; standard code validation against SNOMED CT, ICD-10, ATC, LOINC reference tables. All checks are pandera schemas and pytest suites in version control.

Uniformity. Unit harmonisation with documented conversion factors; category mapping tables to canonical representations; identifier format normalisation. Transformations retain originals and are logged.

Duplicates. Layered: exact duplicates, near-exact (manual review), probabilistic linkage (recordlinkage/Dedupe), and cross-modal verification via genotype concordance (plink) and proteomic intensity correlations.

Missingness. Variable-level completeness stratified by site, personnel, and time. Heatmaps reveal clustered patterns. Design-based skip logic is distinguished from protocol deviations. Threshold-based alerts trigger formal issues.

Plausibility. Clinically defined ranges encoded as pandera schemas; logical consistency rules (age at diagnosis $\leq$ age at last visit; future dates rejected); Interquartile Range / Median Absolute Deviation (IQR/MAD)-based outlier detection per site; heaping and speed-of-entry pattern detection.

Cross-form/cross-modal discordances. Baseline diagnoses vs. problem lists; medication reconciliation across forms; temporal consistency of onset, collection, and withdrawal dates; clinically recorded sex vs. genetically inferred sex.

Site and personnel performance. Completeness rates, edit-check failure rates, query resolution times, and error proxies per site and user, exposed via Grafana dashboards.

1.3 Correction, Cleaning, and Traceability

No value is overwritten silently: every change records original/new values, actor, timestamp, and reason in an append-only audit table. A centralised query log (unique ID, patient/sample/site, severity, assignee, status transitions) drives a structured resolution workflow. Standard Operating Procedures (SOPs) distinguish straightforward corrections, high-impact cases (Principal

Investigator [PI] approval), and sample-level Quality Control (QC) decisions. Cleaned datasets are versioned with data-cut dates and git tags. Quality is reassessed after FHIR and OMOP mapping using validators and the OHDSI DataQualityDashboard.

2 Proposal for Dynamic Analysis

2.1 Architecture

The system separates operational (data capture) and analytical (OMOP CDM) PostgreSQL instances so research queries never degrade clinical workflows. Keycloak tokens carry role and site claims, enforcing governance-aware access with every query logged. All queries, cohort definitions, and report templates are versioned in git.

2.2 Fast Queries

Composite B-tree indexes on OMOP tables (`person_id` + date, `concept_id`), genomic (sample, gene), and proteomic (protein accession) columns. Three nightly-refreshed materialised views: patient-level summary (cross-modal, single-query exploration), longitudinal measurements (time-series trajectories), and site-level completeness (operational dashboards). PgBouncer pools connections; expensive queries route to a streaming read replica.

2.3 Exploration and Correlation

The patient entity links all modalities. A library of parameterised SQL/Python templates encapsulates common cross-modal patterns. Containerised JupyterLab and RStudio Server connect via read-only, role-scoped credentials with pre-installed analytical libraries (pandas, scikit-learn, lifelines, tidyverse, survival, DESeq2, MSstats). OHDSI Atlas provides GUI-based cohort definition exportable as JSON for downstream scripts. Modular workflows support survival analysis, mixed-effects models, clustering, and predictive classifiers with logged metadata.

2.4 Reports and Dashboards

Grafana (operational): enrolment progress, completion rates, missingness heatmaps, query volumes, site scorecards—defined as code, role-scoped. **Apache Superset** (research): interactive charts, ad-hoc filtering, row-level security via Keycloak. **Automated reports**: Airflow/Prefect Directed Acyclic Graphs (DAGs) render parameterised scripts to PDF/HTML via Quarto, tied to script version and data cut.

Adaptability. Schema evolution via Alembic/Flyway; configuration-driven Extract–Transform–Load (ETL) pipelines (JSON/YAML); modular, independently replaceable analytical components; extensible to new modalities (wearables, imaging) and external sources (registries, molecular databases).

2.5 Exploratory Statistical Study

A practical demonstration using synthetic Clinical Data Interchange Standards Consortium (CDISC) data (random.cdisc.data), progressing from demographic summaries through longitudinal trajectories and change-from-baseline analysis to Kaplan–Meier survival curves. The complete interactive report with code and figures is available at **Exploratory Analysis — GitHub Pages**.

3 Mobile App for Prospective Data Collection

3.1 Architecture and Technologies

Seven layers with clear separation of concerns: (1) patient mobile app + clinician portal, (2) API services (FastAPI/NestJS), (3) Identity and Access Management [IAM] (Keycloak—OpenID Connect, RBAC, Multi-Factor Authentication [MFA]), (4) operational PostgreSQL (pseudonymised), (5) de-identification pipelines (Airflow/Prefect), (6) OMOP CDM warehouse (PostgreSQL + OHDSI Atlas), (7) observability (Prometheus, Grafana, Loki, append-only audit log). All FOSS, containerised (Docker/Podman), vendor-independent. The complete system architecture is illustrated in the **Data Flow Diagram (Annex, Fig. 1)**.

The mobile app (Flutter/React Native) is offline-first with encrypted SQLite, background HTTPS sync, save-and-return (critical for DM1 fatigue), large tappable controls, and accessibility features. The clinician portal (React+TypeScript) provides site-scoped data review, annotation, query management, and adherence dashboards. The data manager console configures forms, edit checks, and data locks.

3.2 Data Integrity

FHIR Release 4 (R4) profile validation at ingestion (12 DM1-specific profiles: Patient, RelatedPerson, Encounter, Questionnaire, QuestionnaireResponse, Observation, Condition, MedicationStatement, MedicationRequest, AdverseEvent, Consent, AuditEvent). Edit checks at client and server. Consent verification at every ingestion event. Append-only audit trail. Subject/visit/study-level data locking. Pseudonymisation by design (direct identifiers only in network-segmented Identity Store). Standard terminologies: SNOMED CT, LOINC, ATC, ICD-10, Orphanet (ORPHA:273), OMOP vocabularies via Athena. Concept mapping: WhiteRabbit, Rabbit in a Hat, Usagi.

3.3 Iteration with the Clinical Team

Month 1: paper prototypes with investigators and patients. Month 2: structured walkthroughs in test environment. Month 3: usability sessions with DM1 patients and site staff, findings actioned. Continuous: weekly Clinical Data Management (CDM)–development meetings, bi-weekly data engineering alignment, monthly governance reviews.

4 Regulatory Compliance and Security

4.1 GDPR and EHDS

Lawful basis: explicit electronic Consent (eConsent) (Articles 6(1)(a), 9(2)(a)), plus public health/research bases (9(2)(i), 9(2)(j)). **Data minimisation:** collect only what the protocol requires; optional data require separate consent. **Subject rights:** access, rectification, erasure, and withdrawal supported through documented procedures. **Data Protection Impact Assessment (DPIA)** structured under Article 35. **EHDS alignment:** FHIR + OMOP interoperability, separation of operational/research environments, transparent logging of secondary uses, cross-border safeguards.

4.2 Governance, Traceability, and Access Control

Append-only audit trail recording actor, role, action, object (before/after), UTC timestamp, device/IP, and mandatory reason for corrections. Granular eConsent (core, optional, secondary research, external sharing) with versioned forms and digital signatures. Withdrawal stops collection and flags downstream pipelines.

RBAC enforced via Keycloak + application-level checks across 7 roles and 37 permissions in 11 categories (see the **RBAC Permission Matrix — Annex, Table 1**). Principles: least privilege, site scoping, quarterly review, elevated permissions require approval with expiry.

4.3 Encryption and Incident Response

In transit: Transport Layer Security (TLS) 1.2+ (mutual TLS [mTLS] internal). **At rest:** encrypted PostgreSQL volumes, pgcrypto field-level, encrypted mobile SQLite. Documented incident response: containment, assessment, GDPR notification (Articles 33–34), remediation, review.

4.4 Risk Assessment

Four risks assessed (see the **Risk Likelihood–Impact Matrix — Annex, Fig. 2** and the **full risk narrative**):

Risk	L / I	Mitigation
R1: FHIR/OMOP mapping complexity	H/Maj	Prioritise minimal variable subset; phase broader coverage
R2: Suboptimal DM1 usability	M/Maj	Early, repeated usability testing; accessible design
R3: Slow governance/legal reviews	H/Mod	Start Month 1; named reviewers, agreed turnarounds
R4: FOSS integration complexity	M/Mod	Integration environment from week 1; incremental testing

5 Work Organisation, Prioritisation, and Timeline

Four overlapping workstreams across three months (see the **Project Gantt Chart — Annex, Fig. 3**):

Month 1 — Foundation and Design. Study data specification (all capture elements); version-controlled instruments and schedules; coding strategy (SNOMED CT, LOINC, OMOP mappings); minimal FHIR profile set; draft RBAC matrix; initial governance documents (Data Management Plan [DMP] outline, DPIA structure, consent requirements, data flow diagrams).

Month 2 — Configuration and Implementation. Configured questionnaires, schedules, and edit checks in test environment; mobile/portal prototype walkthroughs; validated FHIR mappings (sample FHIR bundles); OMOP ETL design (WhiteRabbit, Rabbit in a Hat); audit trail and correction workflows; refined DPIA, retention rules, consent language.

Month 3 — Validation and Pilot. DM1 test cases (typical + edge); RBAC and audit trail verification; end-to-end FHIR-to-OMOP validation (Atlas cohort definitions); usability sessions with DM1 patients and staff; finalised DMP and governance docs; pilot scope defined (sites, enrolment targets, monitoring, escalation).

Milestones: Data spec complete (end Mar); test environment configured (end Apr); OMOP ETL validated (mid-May); pilot-ready (end May).

RACI matrix (see the **RACI Responsibility Matrix — Annex, Table 2**): CDM is Responsible/Accountable for most study design, configuration, testing, and documentation tasks. Development/Technical leads FHIR profiles and app prototypes. Data Engineering leads OMOP ETL. Legal/Data Protection Officer (DPO) is accountable for DPIA. Study Leadership for pilot preparation.

Coordination: weekly CDM–Development syncs; bi-weekly CDM–Data Engineering alignment; monthly governance reviews. DM1-specific interface design (simplified screens, proxy mode, save-and-return, configurable reminders) is iterated continuously.

Interoperability targets include European Reference Network for Neuromuscular Diseases (EURO-NMD) registries (Orphanet, Human Phenotype Ontology [HPO]), EHDS secondary use, and CDISC as a secondary mapping target.