Hackathon Requirements Document

Real-Time Operational Dataflow Metrics for Clinical Trials: Integrating Source Data and Presenting for Actionable Insights

Objective

Design and prototype a solution that ingests and integrates the latest available clinical and operational data from multiple sources to generate actionable insights, detect operational bottlenecks, and support scientific decision-making. The solution should leverage Generative and Agentic AI to enhance collaboration between Data Quality Teams (DQT), Clinical Research Associates (CRAs), and Investigational Sites.

Summary of Available Source Data

The solution will use the latest data snapshot from multiple clinical trial systems. Data categories include:

- Patient and Site Operational Metrics: Region, country, site, subject ID, visit/page completion status, subject status.
- Data Quality and Query Metrics: Open queries by type, non-conformant data, query resolution status.
- Lab, Coding, and Safety Reconciliation: Lab issues, coded/uncoded terms, eSAE reviews.
- Form and Verification Status: SDV status, frozen/locked/signed forms, overdue CRFs, inactivated folders.
- Derived Metrics: % missing visits/pages, % clean CRFs, clean patient status.

Feature Engineering Guidance

Key Features to Build

- Data Integration Layer: Unified patient/site-level view from all available data sources.
- Insight Generation: Identify data gaps, unresolved queries, and operational bottlenecks.
- Visualization Dashboard: Real-time drill-down views by region, site, patient, and metric.
- Collaboration Tools: Alerts, tagging, and comments for DQT, CRAs, and sites.

Derived Metrics Summary

Teams must calculate percentages for all key parameters, including missing visits, missing pages, open queries, non-conformant data, and verification status. The clean patient status should be derived as a combination of all these parameters. For example, a patient is “clean” only if there are zero missing visits, no unresolved queries, and all required forms are verified and signed.

Additionally, teams must develop a Data Quality Index that aggregates these metrics into a single score. This index should assign weights to each parameter and reflect the impact of

critical factors (e.g., unresolved safety issues). It should enable rapid assessment of data quality across sites, patients, or trials and support early intervention and decision-making.

Scientific Questions to Address

- Which sites/patients have the most missing visits/pages or unresolved queries?
- Where are the highest rates of non-conformant data?
- Which sites/CRAs are underperforming based on current metrics?
- Where are the most open issues in lab reconciliation or coding?
- Which sites require immediate attention based on current data?
- Can we flag sites with high deviation counts or low query resolution rates?
- Is the current data snapshot clean enough for interim analysis or submission?
- Can readiness checks for statistical deliverables be automated?

Data Sections for Student Use

- Patient Visit Data (visit dates, missing data, adverse events)
- Query Metrics (open/closed queries, resolution time)
- Site Performance (enrollment rate, deviation count)
- CRA Activity Logs (monitoring visits, follow-ups)
- Clean Data Milestones (planned vs. actual cut-off readiness)

AI Capabilities to Explore

- Generative AI: Summarize site performance, generate CRA reports
- Agentic AI: Recommend actions based on current risk signals
- LLM Integration: Natural language querying of snapshot data

Evaluation Criteria

- Innovation: Novel use of AI for snapshot-based insights
- Impact: Potential to reduce delays and improve data quality
- Usability: Intuitive dashboards and workflows
- Scalability: Extendable across studies and therapeutic areas
- Collaboration: Supports cross-functional teamwork

Deliverables

- Working prototype or mock-up
- Presentation deck explaining the solution
- Documentation of architecture and AI models used
- Sample data used (if simulated)