Project Report: Clinical Trial Patient Recruitment and Adherence Monitoring

1. Executive Summary and Project Objectives

This project successfully implemented a **central command center** to monitor the performance of a multi-site clinical trial. The primary objective was to address the immense costs and delays associated with poor patient recruitment and non-adherence by providing trial managers with **real-time, actionable insights**.
The solution integrates clinical trial data (screening, enrollment, visits) with predictive modeling to create a secure, interactive Business Intelligence (BI) dashboard. The key deliverables include visualizing the recruitment funnel, assessing site performance, tracking patient adherence, and, critically, **flagging patients at high risk of dropout** using a machine learning model.

2. Data Engineering and Data Modeling

2.1. Data Sources and Integration

The solution was built upon the integration of five key simulated data extracts from the Electronic Data Capture (EDC) system, which were loaded and processed for analysis. These files served as the foundation for the Star Schema design:

File Name	Description	Key Columns Used for Analysis
site_metadata.csv	Details on sites, including country and target enrollment.	site_id, country, target_enrollment
screening.csv	Records of patient screening attempts and failure reasons.	date_screened, screen_status, failure_reason
enrollment.csv	Patient enrollment and randomization status and dates.	date_enrolled, randomized, date_randomized
visits.csv	Detailed visit history, adherence, and diary submission.	visit_status, medication_adherence_pct, diary_submitted
UserPatientMap.csv	Mapping file used to enforce Row-Level Security (RLS).	patient_id, usermail

2.2. Data Transformation and Preparation

The data integration pipeline focused on cleaning and preparing the data for both the predictive model and the Power BI reports. Key transformations included:

- **Pseudonymization:** Clinical identifiers were kept secure, and a separate mapping file (UserPatientMap.csv) was utilized to link non-sensitive user emails to patient IDs for RLS purposes.
- **Date Handling:** All date columns (e.g., date_enrolled, scheduled_date, actual_date) were converted to the correct data type.
- **Feature Engineering for Modeling:** The features critical for the dropout model were engineered, such as:
    - Calculating **Missed Visit Count** per patient (derived from visits.csv).
    - Calculating **Average Medication Adherence** per patient.
    - Calculating **Total Observed Days** in the trial.

2.3. Data Model (Star Schema)

A robust **Star Schema** was constructed in Power BI (as implemented in InfotactProject1.pbix) to ensure fast and scalable reporting.

- **Fact Table:** The main visits data served as the core Fact table, capturing events and key metrics.
- **Dimension Tables:** site_metadata and transformed patient tables served as Dimension tables, allowing for slicing and dicing the metrics by site, country, and patient characteristics.

3. Dropout Risk Prediction Model Development

A central component of this project was the development and integration of a Machine Learning model to forecast patient dropout risk.

3.1. Model Selection and Training

- **Model Type:** A **Classification Model** - Logistic Regression was trained using Python/JupyterNotebook/Google Colab). The model's goal was to predict a binary outcome: Dropout (1) or No Dropout (0).
- **Feature Importance:** The model was trained on engineered features that are highly predictive of adherence, such as the total count of **missed visits**, low **medication adherence percentages**, and patterns of **rescheduled visits**.

3.2. Model Integration and Output

- **Scoring:** The trained model was used to score the current patient population, generating the **dropout_predictions.csv** file. This file contains the patient_id and the calculated binary risk prediction (0 or 1).
- **BI Integration:** This prediction file was imported directly into the Power BI data model. This allowed for the creation of a powerful **"At-Risk" patient visual** on the dashboard, where patients were clearly labeled as "Low Risk" or "At Risk" (as seen in the PDF export).

Key Finding: The model identified a significant portion of the currently enrolled patients as **"At Risk,"** with the dashboard indicating 43.04% **of the patients in the trial are flagged as high risk**, necessitating immediate intervention by site monitors.

4. Business Intelligence and Actionable Insights

The Power BI dashboard was designed as a "central command center," providing granular insights across the three main project requirements: Recruitment, Site Performance, and Adherence.

4.1. Recruitment and Enrollment Funnel

The dashboard provides a clear visualization of the patient recruitment pipeline:

- **Screened vs. Enrolled vs. Randomized:** The funnel clearly tracks patient flow, with an observed **Screen Failure Rate of** 0.21.
    - Example Metrics from PDF: Total Screened (100), Total Enrolled (17), Total Randomized (62).
- **Recruitment Bottlenecks:** The screening.csv data enabled analysis of common **Screen Failure Reasons** (e.g., 'Low Hb', 'Non-Compliance', 'Medical History'), allowing trial managers to refine pre-screening criteria or site training.

4.2. Patient Adherence and Visit Monitoring

The dashboard provides deep insights into patient compliance:

- **Visit Status Breakdown:** A crucial visual breaks down total patient visits by status: **Completed, Missed, and Rescheduled**. The high percentage of completed visits (83.54%) is positive, but the presence of Missed and Rescheduled visits (12.03% and 4.43% respectively) highlights the need for proactive monitoring.
- **Medication Adherence:** The **Average Medication Adherence** KPI allows for a quick assessment of compliance across the study and per site.

4.3. Site Performance Leaderboard

Site performance metrics enable the identification of best practices and underperforming sites:

- **Enrollment Velocity:** Sites were ranked by the number of **Enrollments**. For example, **SITE05** was identified as the top-performing site for enrollment.
- **Adherence Leaderboard:** The dashboard compares the **Average Medication Adherence** score for each site (e.g., SITE05 had the highest adherence at 76.98%, while SITE01 had the lowest at 56.52%). This pinpoints sites that may require additional support or training on patient management.

5. Security and Conclusion

5.1. Row-Level Security (RLS) Implementation

Security was paramount, particularly in a system dealing with patient data. **Row-Level Security (RLS)** was implemented in Power BI using the **UserPatientMap.csv** file.

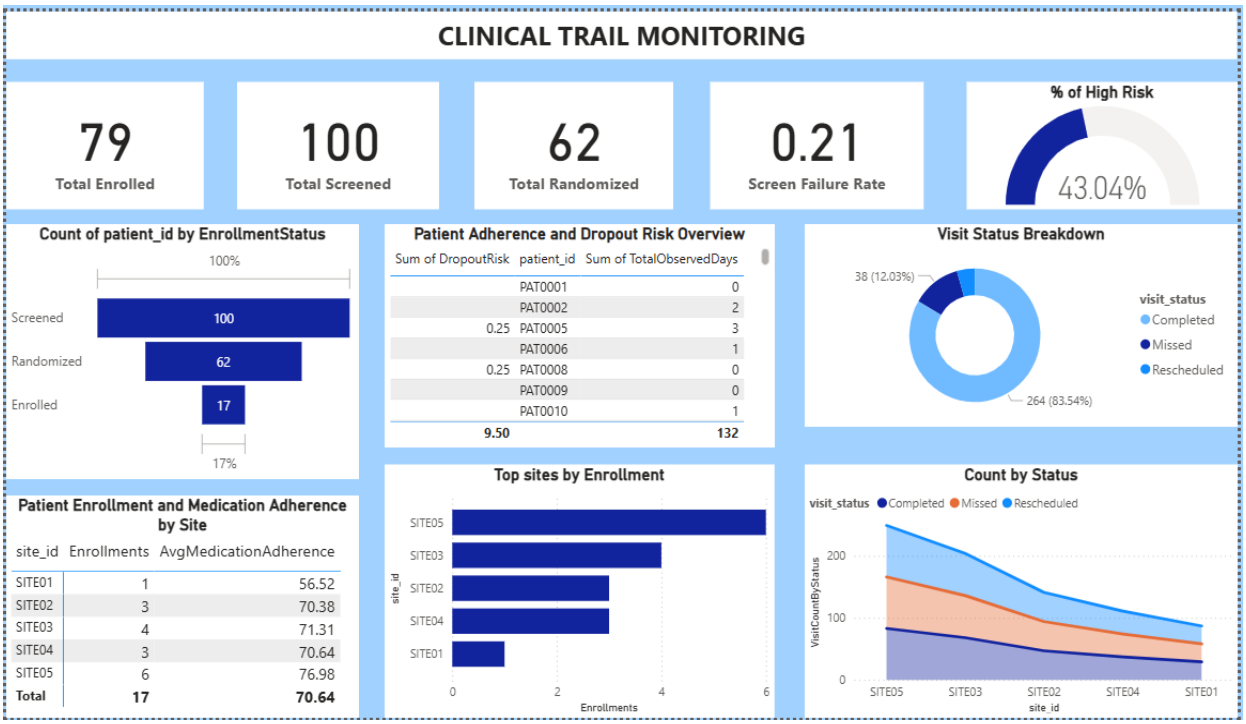
- **Mechanism:** RLS ensures that site monitors or regional managers accessing the Power BI report can **only view data for the patients/sites they are directly responsible for**. This strictly adheres to minimal data access principles, enhancing data privacy and compliance.

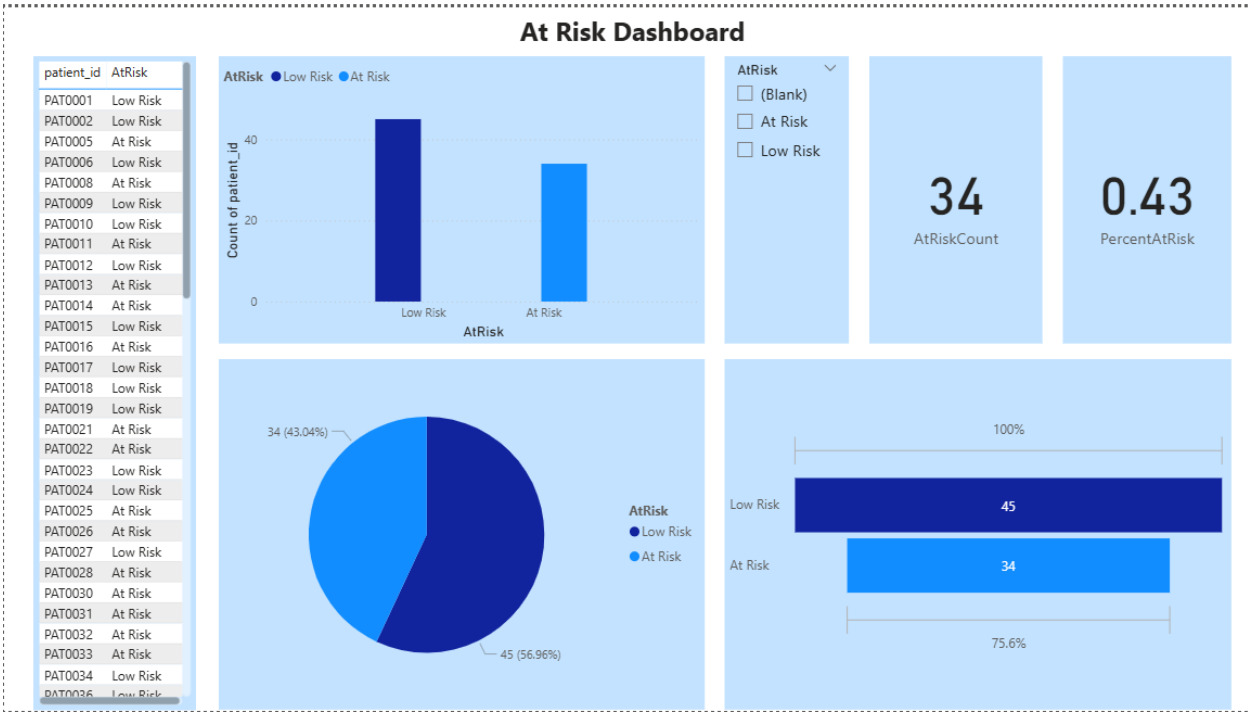
5.2. Conclusion and Future Enhancements

The Clinical Trial Monitoring Command Center provides a powerful, data-driven tool to improve trial efficiency and patient retention. By combining historical data analysis with predictive risk scores, trial managers can shift from reactive troubleshooting to **proactive patient intervention**.

Suggested Future Extensions:

1. **Alert Automation:** Integrate the dropout risk score with an external system (e.g., via a cloud function) to automatically send **SMS/email alerts** to site coordinators when a patient's risk score crosses the predefined threshold.
2. **Survival Analysis:** Implement an advanced **Survival Analysis** model to not only predict *if* a patient will drop out but also estimate the **time-to-dropout**, enabling better resource planning and intervention timing.
3. **Explainability:** Incorporate feature-level explainability (e.g., pre-calculated SHAP values) into the dashboard to show *why* a patient was flagged (e.g., "Flagged primarily due to 3 Missed Visits and Adherence below 60%").





Date: 05 October 2025
Author: Shriraksha P Acharya