DATA QUALITY ASSESSMENT REPORT

PREPARED BY: EMMANUEL KWABENA ANSU (JUNIOR DEVELOPER)

MEDTRACK GHANA-PATIENT APPOINTMENT DATASET

Date: 2nd February 2026

EXECUTIVE SUMMARY:

Analysis of the sample CSV revealed violation in all six data quality dimensions. The most severe issues are duplicate Patient IDs, inconsistent phone number and date formats, missing patient names, and inconsistent doctor/payment representations. These directly cause the reported symptoms: failed SMS reminders, inflated patient counts in reports and incorrect billing.

DATASET FOR REFRENCE:

Patient ID	Patient Name	Phone Number	Appointment Date	Doctor Name	Payment Status
P001	Kwame Mensah	0244123456	2025-10-15	Dr. Osei	paid
P002	Ama Serwa	244789012	15/10/2025	dr. osei	Paid
P001	Kwame Mensah	0244123456	2025-10-20	Dr. Adjei	Pending
P003	Kofi Annan	0244567890	10/16/2025	Dr. Osei	Failed
P004	(missing name)	0555234567	2025-10-17	Dr. Mensah	Paid
P002	Ama Serwa	0244789012	2025-10-15	Dr. Osei	paid

TASK 1.

DATA QUALITY DIMENSION	VIOLATION EXAMPLE FROM DATASET	EXPLANATION
Accuracy	**Phone Number**: 244789012 (P002)	Missing leading zero. All phone numbers must start with 0.
Completeness	**Patient Name:** (empty) for P004	Patient Name is a required attribute for identification, reporting and clinical workflows
Consistency	Doctor's Name written as Dr.Osei & dr.osei	Same doctor written in different ways (capitalization, title) causing duplicate doctor counts by breaking assumptions in the system.
Timeliness	**Ambiguous date format**: 2025-10-15, 15/10/2025 & 10/16/2025	Mixed formats cause sorting/filtering errors and confusion in reminder systems
Validity	**Appointment Date:** 15/10/2025 & 10/16/25 of P002 and P003	Both appointment dates deviate from the default date format of YYYY-MM-DD to completely different date formats
Uniqueness	**PatientID** P001 appears twice. P002 also appears twice with different phone numbers	Same patient duplicated which causes double counting in reports.

TASK 2: BUSINESS IMPACT OF EACH ISSUE

Below shows how each issue leads to real operation problems.

ISSUE	OPERATIONAL PROBLEMS	BUSINESS FUNCTION MOST AFFECTED
Inaccuracy: Incorrect Phone number format	• SMS appointment reminders fail because telecom systems require valid phone number formats. • Patients do not receive notifications for appointments, test results, or schedule changes.	• **Operations** (patient attendance, scheduling efficiency)
Duplicate Patient Records (Uniqueness)	• Reports overcount patients. • Billing may be attempted multiple times or skipped due to ambiguity. • Clinical staff may see fragmented appointment histories	• **Finance** (billing failures, revenue leakage) • **Clinical** (incomplete patient history)
Inconsistent Date Formats (Consistency & Timeliness)	• Reminder jobs fail or run at incorrect times. • Appointments may appear overdue or upcoming incorrectly in dashboards.	• **Operations** (appointment management, reminders)

Missing Patient Name (Completeness)	• SMS personalization fails (“Dear, your appointment…”). • Staff cannot confidently confirm patient identity.	• **Clinical** and **Operations**
Uniqueness Issue	• Inflated patient metrics • Misleading KPIs for management • Poor strategic decisions (staffing, capacity planning)	• **Operations** • **Management / Analytics**

TASK 3 RCOMMENDED SOLUTIONS (TOP 3 CRITICAL ISSUES)

Rank	Issue	Technical solution	Who should implement	How to verify
1	Inaccurate phone numbers	Implement phone numbers validation rules during data entry. It must start with 0 and must be exactly 10 digits	Backend engineers (validation rules)	Send test batch SMS messages to a subset and confirm successful delivery from the SMS gateway logs
2	Duplicate Patient Records	Enforce unique PatientID constraint in the database. Add a duplicate-detection script that checks for duplication	Database administrator (DBA)	Run patient count reports before and after cleanup and confirm duplicates drop to zero.
3	Invalid Date formats	Standardize all dates to a single format. Use a date parsing script that converts know acceptable formats and flags invalid one. Additionally, add frontend/backend validation to block dates formats not defined.	Backend engineers	Attempt to import mixed-format test data. Confirm appointment reminders and billing triggers run with correct schedules.

TASK 4

In the field of cybersecurity, poor data consistency in patient and staff records creates a dangerous security gap. When identifiers such as name, IDs, phone numbers, emails, doctor names) appear in different formats, the system cannot reliably determine who is who. This undermines identity integrity, the foundation of all cybersecurity controls and exposes the organization to fraud, unauthorized access and compliance violations.

Identity Spoofing & Access Control Bypass: Poor data consistency creates a dangerous opening in identity management, especially when user or patient identifiers are stored with variations such as “Dr. Osei,” “dr. osei,” or “Dr.Osei.” When the system interprets these inconsistencies as separate entities rather than one individual, it weakens the foundational logic behind access control. In a healthcare environment, this flaw can be intentionally exploited: an attacker or malicious insider may create a slightly altered version of an existing identity to bypass access restrictions or escalate privileges. Because the system cannot reliably match these variations to a single authentic user, unauthorized individuals may gain access to medical records, back-office systems, or administrative capabilities. This is critical in healthcare, where identity-based controls are the primary safeguard for determining who is allowed to view, edit, or transmit sensitive patient information. Once consistency breaks down, the security model collapses, and access control becomes unreliable, exposing patient data to breach, manipulation, or misuse.

Audit Trail Fragmentation: Audit trails are a core security requirement in healthcare, and their reliability depends entirely on the system’s ability to tie all activity back to one correctly identified user. When identifiers are inconsistent, the logs fragment into multiple partial identities, each recording only a slice of actual behaviour. This fragmentation makes suspicious activities appear normal because critical actions such as unauthorized edits, unusual login attempts, or abnormal access patterns are distributed across multiple identity variants. As a result, investigators cannot

reconstruct timelines or identify connections between related events. When logs fail to correlate, incident response teams lose visibility, undermining the ability to detect tampering, policy violations, or breaches of protected health information (PHI). Ultimately, malicious actions can occur without triggering alarms, and even when discovered later, the lack of a coherent audit trail makes forensic investigation nearly impossible.

Fraud, Duplicate Billing & Ghost Records: Data inconsistency in patient records directly enables financial and operational fraud within healthcare systems. When names, IDs, or contact information vary even slightly, the system may accept multiple versions of the same patient as separate individuals. This allows fraudulent actors to create duplicate or "ghost" patient profiles that can be used to submit false claims, schedule appointments that never happened, or manipulate billing workflows without detection. Because the system cannot distinguish between legitimate and artificially duplicated records, attackers can hide fraudulent transactions within the noise of inconsistencies. This creates serious financial risk, complicates revenue-cycle management, and undermines the accuracy of patient histories. Ultimately, inconsistent identifiers make it far easier for fraud to blend into everyday processes, enabling attackers to exploit gaps without triggering financial controls or audit safeguards.

Security Monitoring Blind Spots: Security monitoring systems such as SIEM platforms, anomaly-detection engines, and fraud-detection algorithms rely entirely on clean, normalized data to accurately identify patterns, detect threats, and correlate events. Inconsistent identifiers break this analytical foundation by preventing the monitoring tools from linking related activities across systems. When identifiers differ in format, casing, or structure, the monitoring engine interprets them as separate entities, resulting in uncorrelated events and incomplete threat patterns. This leads to missed alerts, false negatives, and an overall degradation in detection capability. Suspicious behaviour appears harmless because the system cannot see the full picture, allowing attackers to operate undetected. In this environment, legitimate threats fade into background noise, leaving the organization exposed to breaches, insider misuse, and long-term compromise of sensitive healthcare data.