International Journal For Research In Advanced Computer Science And Engineering

ISSN: 2208-2107

Volume 11 Issue 01 March 2025

Available on-line at:https://ijracse.org/index.php/cse/article/view/2474/1384

DOI:https://doi.org/10.53555/cse.v11i1.2474 *Received: 07-Dec-2024 Accepted: 10-Jan-2025*

A FRAMEWORK FOR ENSURING DATA INTEGRITY IN HEALTHCARE DATA MIGRATION

Jagrutiben Padhiyar[1*], Akash Narendrakumar Parmar[2]

[1]Senior Application Developer Gujarat Technological University (Bachelor of Engineering in Information Technology) jagrutipadhiyar6@gmail.com*

[2]Rajiv Gandhi Proudyogiki Vishwavidyalaya (R.G.P.V) Bachelor in Mechanical Engineering
kash.parmar13@gmail.com

Corresponding Author:
***Email:** jagrutipadhiyar6@gmail.com*

Abstract

The digital transformation of healthcare systems has accelerated the adoption of electronic health records (EHRs) and data-driven technologies. As healthcare institu- tions transition from legacy systems to modern platforms, data migration has become a critical process for consolidating, storing, and analyzing patient data. Despite its im- portance, data migration introduces significant risks to data integrity. Missing records, corrupted values, duplication, and schema mismatches during migration can lead to inaccurate or incomplete datasets. In the healthcare domain, such inconsistencies can have severe consequences, including misdiagnosis, delayed treatments, financial errors, and regulatory non-compliance. Therefore, maintaining strict data integrity during migration is essential to ensuring patient safety, trustworthiness of clinical decisions, and adherence to standards such as HIPAA, GDPR, and HL7/FHIR.

Healthcare data is highly complex, often containing patient demographics, diag- nostic information, laboratory results, and treatment histories. The heterogeneity of formats—ranging from spreadsheets and CSV files to JSON structures and relational SQL databases—compounds the challenge of reliable migration. Traditional migration tools primarily

focus on transferring data efficiently but lack comprehensive mechanisms for validating integrity at every stage. This gap necessitates a structured framework that integrates validation techniques to guarantee that data remains accurate, complete, and consistent after migration.

This study addresses these challenges by proposing a framework for healthcare data migration that emphasizes integrity validation. The methodology involves simulating the migration of an open healthcare dataset from flat file formats (CSV/Excel) into JSON and SQL databases. Pre- and post-migration validation is performed using au- tomated methods such as record count comparison, duplicate detection, and checksum verification. Additionally, domain-specific data quality rules are applied to verify the accuracy of mandatory attributes such as patient ID, age, and diagnosis. Python li- braries, including pandas for data manipulation and hashlib for checksum computation, are employed to design a lightweight validation pipeline. The framework aims to provide a reproducible and efficient approach to healthcare data migration, with the ultimate goal of minimizing risks and ensuring data reliability in clinical environments.

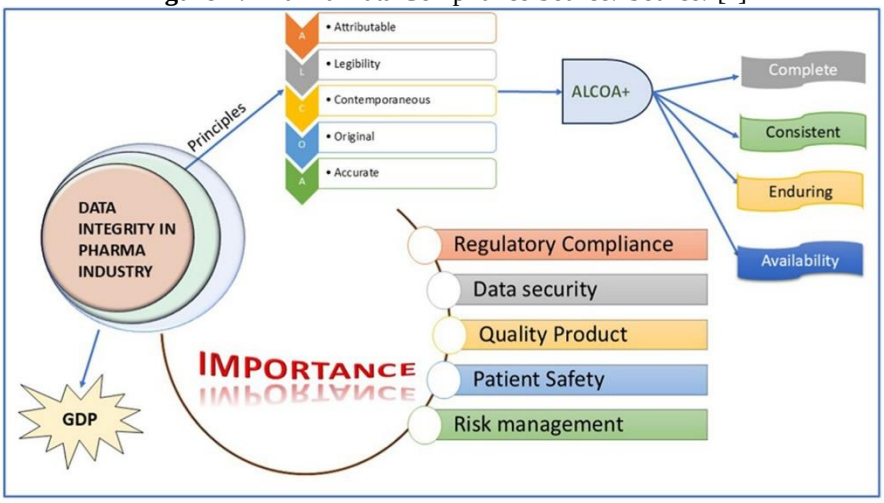
Keywords: *Data Migration, Data Integrity, JSON, SQL*

1 Introduction

The healthcare sector is witnessing a paradigm shift towards digitization, driven by the rapid adoption of electronic health records (EHRs), cloud storage platforms, and advanced analytics systems. Data migration, which involves transferring healthcare information from one system or format to another, plays a critical role in this transformation. Migration may occur when healthcare organizations adopt a new information system, consolidate multiple sources of data, or transition to cloud-based architectures. The process is not merely a technical necessity but a strategic requirement for ensuring that historical and real-time patient information remains accessible, usable, and secure.

While data migration promises improved efficiency and scalability, it introduces significant risks to data integrity. Data integrity refers to the accuracy, consistency, and reliability of data across its lifecycle. In healthcare, the consequences of compromised integrity are particularly severe. For example, if patient identifiers are lost or mismatched during migration, clinical staff may fail to retrieve correct records, leading to misdiagnosis or incorrect treatments. Similarly, corrupted or missing diagnostic information could affect research quality, while duplicate entries can distort statistical analyses and billing systems. Beyond clinical outcomes, compromised data integrity also poses risks to regulatory compliance, as healthcare providers are required to adhere to standards such as HIPAA (Health Insurance Portability and Accountability Act), GDPR (General Data Protection Regulation), and HL7/FHIR (Fast Healthcare Interoperability Resources).

Figure 1: Pharma Data Compliance Source: Source: [1]



Ensuring data integrity during migration is challenging due to the heterogeneous nature of healthcare data. Healthcare information exists in various formats, such as spreadsheets (CSV/Ex- cel), semi-structured documents (JSON, XML), and structured relational databases (SQL, Oracle). Additionally, data is often distributed across multiple legacy systems with inconsistent schemas, coding systems, and validation rules. Traditional migration approaches primarily emphasize efficiency—moving large volumes of data quickly—without sufficient emphasis on validating accuracy and completeness at each stage. As a result, healthcare organizations face increased vulnerability to undetected errors and inconsistencies post-migration.

1.1 Background and Problem Statement

Healthcare data is inherently diverse, encompassing demographic details, clinical histories, diag- nostic reports, and laboratory results. This diversity leads to the coexistence of multiple storage formats within a single healthcare ecosystem. For example, older systems may store data in CSV or Excel files, while modern systems adopt relational databases or JSON-based APIs. Migrating data across these formats is far from trivial. Even minor inconsistencies—such as missing patient IDs, incorrect age entries, or incomplete diagnosis codes—can compromise the trustworthiness of migrated datasets.

The complexity is further heightened by the volume of data generated daily in healthcare environments. Manual validation of migrated records is impractical, leaving organizations vul- nerable to undetected data quality issues. Current migration strategies often prioritize speed and functionality, while validation is treated as an afterthought, performed only when errors are visible. This reactive approach introduces risks of data corruption, duplication, or loss that undermine the reliability of healthcare information systems. The lack of a systematic framework that addresses both the technical and domain-specific challenges of healthcare data migration represents a significant gap in current research and practice.

1.2 Objectives of the Study

This research seeks to address the above gap by proposing a framework for ensuring data integrity in healthcare data migration. The study is designed around the following objectives:

- To simulate the migration of a publicly available healthcare dataset across heterogeneous formats, specifically from flat files (CSV/Excel) into JSON and relational database structures.
- To design and implement a set of automated validation techniques, including record count comparisons, duplicate detection, and checksum verification, that ensure accuracy, consis- tency, and completeness during migration.
- To define healthcare-specific quality rules, focusing on critical attributes such as patient ID, age, and diagnosis, ensuring that essential information is not lost or altered during migration.
- To implement a lightweight Python-based validation pipeline using libraries such as pandas and hashlib, providing a reproducible mechanism for anomaly detection and reporting.

These objectives aim to bridge the gap between technical efficiency and clinical reliability, demonstrating that data migration can be both fast and trustworthy when supported by a struc- tured validation framework.

1.3 Contribution and Scope

The contribution of this research lies in developing and testing a reproducible framework that integrates validation as a central component of the data migration process. Specifically, this study contributes by:

- Designing a Python-based validation pipeline that ensures record-level integrity across mi- gration formats.
- Embedding domain-specific quality rules tailored for healthcare data, addressing the unique risks of this sector.
- Demonstrating the practical applicability of the framework through experiments on publicly accessible healthcare datasets, thereby showing its relevance in real-world scenarios.

The scope of this research is intentionally defined to focus on open healthcare datasets and simulated migration scenarios. While the experimental setup does not replicate the full scale and complexity of enterprise healthcare systems, the findings offer valuable insights into the preservation of data integrity in smaller, controlled settings. The framework presented here is scalable and adaptable, providing a foundation for future implementation in clinical environments where compliance with standards such as HIPAA and HL7/FHIR is mandatory.

2 Literature Review

Such research by[7] focuses on the challenges of migrating electronic health records (EHRs) from legacy systems to modern healthcare platforms. The study emphasizes that interoperability across different systems is often hindered by heterogeneous data formats, including CSV, Excel, relational databases, and semi-structured JSON. Their research highlights that improper migration can lead to missing records, duplicated patient information, or corrupted fields, all of which can affect clinical decision-making. The authors conducted a case study in multiple hospitals to evaluate the effectiveness of existing migration frameworks. They observed that many institutions rely heavily on ETL pipelines that prioritize speed over validation. Such pipelines frequently overlook domain-specific validation rules, such as verifying mandatory attributes like patient ID or diagnosis codes. Kuo and Kushniruk also stress that regulatory compliance, including HIPAA and HL7 standards, necessitates systematic verification during migration. Their findings suggest that migration processes must integrate integrity checks at every stage to prevent operational and clinical risks. The research provides a foundation for designing validation frameworks that are both technical and domain-aware. It also emphasizes the need for reproducible and scalable methods suitable for large datasets. The study's methodology combines empirical evaluation with system analysis, offering insights into the limitations of existing tools. Overall, this work demonstrates the importance of structured, validated migration processes in healthcare environments.

Several studies by [5] examine data migration as an essential part of system evolution, highlight- ing the risks of information loss during system upgrades. Their research identifies that incomplete or improper migration can compromise institutional reliability and decision-making processes. The authors analyze historical case studies where legacy healthcare systems were decommissioned, noting that critical patient information was sometimes corrupted or lost. They propose a systematic approach to migration that emphasizes consistency, accuracy, and completeness of data. According to their findings, most organizations underestimate the complexity of migration and assume that data transfer alone ensures correctness. Such assumptions often result in duplicated records, missing attributes, and inconsistencies in patient histories. Bisbal et al. highlight that validation is not merely a technical issue but also a strategic concern affecting clinical operations. They recom- mend integrating validation techniques, including checksum verification and duplicate detection, into migration workflows. Their study provides empirical evidence that structured frameworks can prevent many common errors in healthcare data transfer. Additionally, they advocate for documenting migration processes to ensure reproducibility. The research also underscores the need for domain-specific rules in healthcare, such as verifying mandatory patient identifiers and diagnostic codes. Finally, the study concludes that successful migration requires a combination of

technical precision, procedural rigor, and continuous monitoring.

Such research by [3] explores data quality and integrity in information systems, offering method- ologies that can be applied to healthcare data migration. The authors focus on validating datasets through record-level checks, duplicate detection, and schema comparisons. Their study argues that ensuring completeness and consistency is critical for high-stakes domains like healthcare, where errors can affect patient safety. They present a comprehensive framework for data quality assessment, including tools to identify missing values, outliers, and inconsistent entries. Batini and Scannapieco highlight that conventional migration approaches often neglect these aspects, focusing primarily on data transformation and loading efficiency. The research emphasizes the importance of automated validation pipelines to manage large-scale datasets. They also stress that domain-specific rules, such as verifying age, diagnosis, and treatment attributes, are essential for maintaining data reliability. The study provides a set of best practices for implementing data quality checks systematically during migration. Their methodology is supported by both theoretical frameworks and practical case studies. The research underscores that combining technical valida- tion with domain-specific rules ensures higher data integrity. Furthermore, the study highlights the benefits of reproducibility and documentation in maintaining long-term reliability. Overall, this research establishes that robust validation mechanisms are central to safe and effective healthcare data migration.

Several studies by [13] focus on schema mapping and transformation as integral components of data migration. They investigate how heterogeneous systems, including relational databases and semi-structured formats, can maintain consistency during data transfer. Their research demon- strates that improper schema alignment often leads to data corruption, loss of meaning, or mis- matched records. Rahm and Bernstein propose a systematic approach to schema mapping that includes automated detection of attribute correspondences and semantic validation. The study emphasizes that syntactic correctness alone is insufficient, as semantic errors may persist even when records appear accurate. Their findings highlight the importance of combining schema mapping with integrity validation techniques, such as checksum verification and mandatory attribute checks. Additionally, they note that large-scale datasets, common in healthcare, require automated tools to perform these validations efficiently. The research provides evidence that structured schema transformation can significantly reduce migration errors. Rahm and Bernstein also discuss the role of reproducibility and auditing in ensuring migration reliability. They advocate for frameworks that integrate schema mapping, data quality assessment, and domain-specific validation rules. Their study concludes that comprehensive approaches combining technical, semantic, and procedural checks are necessary for effective healthcare data migration.

Such research by [8] investigates the use of cryptographic hash functions for verifying data integrity during migration. The authors focus on MD5 and SHA-256 hashing as methods to detect corruption or tampering in large datasets. Their study shows that checksum validation can identify even minor alterations that would otherwise remain undetected. Li et al. highlight that while hashing is effective for detecting changes at the byte level, it does not guarantee semantic correctness or the presence of mandatory fields. The research emphasizes that checksum techniques should be combined with other validation methods to ensure comprehensive integrity. In health- care, for instance, hash validation must be supplemented by checks on patient ID, age, diagnosis, and treatment attributes. The study also provides practical guidance on implementing automated pipelines using hashing, especially for datasets transferred across formats such as CSV, JSON, and SQL databases. The authors demonstrate that combining checksum validation with record count comparisons and duplicate detection increases reliability. Their methodology is particularly useful for environments where large volumes of data are migrated frequently. Li et al. also stress the importance of documenting validation results for auditing and regulatory compliance. Overall, the research provides a foundational technique for automated integrity verification in healthcare data migration. Several studies by [1] examine the challenges of migrating healthcare data to cloud- based platforms. Their research identifies scalability, security, and interoperability as key concerns that must be addressed during migration. Al-Hamdani et al. observe that cloud systems introduce additional complexity due to distributed storage and varied data formats. Their study emphasizes the importance of integrating automated validation pipelines that include checks for missing records, duplicates, and mandatory attributes. They also note that cloud-based migration requires continuous monitoring to detect errors in real time. The research highlights that standard ETL processes often fail to capture domain-specific validation needs, leaving clinical data vulnerable. Additionally, they suggest that combining hashing, record comparison, and semantic validation can improve overall data integrity. The study provides case studies demonstrating the effectiveness of integrated validation in reducing errors. Their findings underline the necessity of reproducible and transparent frameworks. Al-Hamdani et al. conclude that healthcare organizations must prioritize both performance and integrity during cloud migrations. Finally, the research establishes guidelines for designing automated, scalable, and compliant migration workflows.

Such research by [16] explores the impact of semantic validation on healthcare data integrity. Their study differentiates between syntactic correctness, which ensures that data formats match, and semantic correctness, which ensures that data retains clinical meaning. Weiskopf and Weng highlight that errors in semantic mapping can occur even when syntactic validation passes. For example, mismatches in ICD-9 and ICD-10 codes can lead to clinically inaccurate datasets de-spite appearing technically valid. Their research demonstrates that semantic inconsistencies can significantly affect clinical decisions, research outcomes, and billing processes. They advocate for frameworks that integrate semantic

checks alongside conventional integrity validations. The study provides methodologies for detecting and correcting semantic errors during migration. It also emphasizes the importance of domain-specific validation rules, including checks on patient demographics, diagnosis codes, and treatment histories. Weiskopf and Weng's findings support the integration of semantic, syntactic, and procedural checks in automated pipelines. Their research contributes to understanding how domain-specific quality rules enhance data reliability. Addition- ally, they highlight the need for documentation and reproducibility in validation frameworks. The study concludes that semantic validation is essential for maintaining the usefulness of healthcare data post-migration.

Several studies by [15] focus on the broader framework of data quality, identifying dimensions such as accuracy, completeness, consistency, and timeliness. Their research argues that quality must be assessed at both technical and domain-specific levels. In healthcare, incomplete or inaccurate data can lead to misdiagnosis, incorrect treatments, and research errors. Wang and Strong propose methodologies for systematically assessing data quality during migration, including automated validation pipelines. They emphasize that maintaining data integrity requires moni- toring at multiple stages, from extraction to loading. Their study demonstrates that combining technical checks, such as record counts and checksum verification, with domain-specific rules im- proves reliability. They also highlight that documenting validation results enhances accountability and supports regulatory compliance. The research provides guidelines for designing reproducible frameworks suitable for large-scale healthcare datasets. Wang and Strong argue that integrating quality assessment into migration processes is essential for operational and clinical safety. Their work underlines the importance of standardizing validation practices across institutions. The study concludes that data quality frameworks are fundamental to successful healthcare data migration. Such research by [4] examines automated pipelines for detecting anomalies and duplicates in large datasets. Their study focuses on optimizing validation processes to handle the scale and complexity of healthcare data. Batini et al. emphasize that manual verification is impractical due to the volume of records, necessitating automated solutions. They describe techniques for detecting duplicates, missing values, and inconsistencies across multiple data formats. Their research demonstrates that integrating multiple validation methods improves overall data integrity. The study also stresses the need for domain-specific rules, such as verifying patient identifiers and treatment codes. Batini et al. highlight the importance of maintaining reproducibility and transparency in automated frameworks. Their findings show that automated validation can significantly reduce errors while increasing efficiency. Additionally, they provide case studies illustrating practical implementations in healthcare settings. The research concludes that scalable, automated frameworks are essential for modern data migration challenges. Several studies by [12] examine schema evolution and its impact on data integrity during migration. Their research emphasizes that changes in data models can create inconsistencies if not properly managed. They propose automated schema matching and semantic validation techniques to ensure that data retains meaning after migration. The study highlights that healthcare systems frequently evolve, requiring adaptive validation mechanisms. Rahm et al. demonstrate that combining schema checks with record-level validation and checksum techniques reduces migration errors. They also underscore the importance of domain-specific validation rules for healthcare attributes. The research provides practical methodologies for implementing automated, reproducible frameworks. Their findings show that integrated validation improves both technical correctness and clinical reliability. Rahm et al. advocate for continuous monitoring and documentation during migration processes. They conclude that schema-aware validation frameworks are critical for maintaining data integrity in evolving healthcare environments.
Such research by [2] investigates the challenges of ensuring integrity in cloud-based healthcare data migration pipelines. Their study identifies distributed storage and concurrent access as factors that increase error risk. They propose a framework that combines record counts, duplicate detection, checksum validation, and healthcare-specific rules. The study demonstrates that an integrated approach significantly improves data reliability. Al-Hamdani et al. also explore the role of automated notifications and logging for error tracking. Their research highlights the importance of reproducibility and transparency in validation frameworks. They argue that cloud migration introduces unique challenges that require both technical and domain-specific checks. The study provides guidelines for designing lightweight, scalable pipelines. It emphasizes that combining multiple validation techniques enhances overall integrity. Al-Hamdani et al. conclude that healthcare organizations must adopt integrated validation to maintain compliance and safety. The research contributes to establishing best practices for modern, cloud-based migration scenarios. Several studies by [9] focus on machine learning approaches for anomaly detection in healthcare data migration. Their research shows that AI-based techniques can complement rule-based valida- tion by identifying subtle errors that may go undetected by conventional methods. Li et al. propose algorithms to detect inconsistencies in patient records, duplicate entries, and semantic mismatches. Their study emphasizes that machine learning can improve scalability and efficiency of validation pipelines. They also highlight that AI methods must be trained with domain-specific constraints to ensure clinical relevance. The research provides experimental evidence that combining AI with traditional validation methods enhances overall integrity. Li et al. argue that automated frameworks integrating multiple approaches are essential for large, heterogeneous datasets. Their findings underscore the potential for advanced tools to support healthcare data reliability. The study also stresses documentation and reproducibility as critical components of any AI-assisted

validation pipeline. Finally, Li et al. conclude that integrating AI into migration frameworks can address limitations of conventional rule-based approaches while maintaining safety and accuracy.

Such research by[10] emphasizes the importance of participatory design principles in defining data migration processes within healthcare institutions. The study advocates for involving end- users—such as clinicians, data analysts, and IT staff—in the planning and execution phases of data migration. By incorporating their insights, the research aims to develop migration strategies that are not only technically sound but also aligned with the practical needs and workflows of healthcare professionals. MacKenzie identifies several challenges, including resistance to change, lack of training, and concerns about data integrity, which can hinder successful migration efforts. The study suggests that addressing these issues through collaborative design can lead to more effective and accepted migration processes. Furthermore, the research highlights the necessity of continuous feedback loops during the migration process to ensure that the system meets the evolving needs of its users. MacKenzie proposes a framework that integrates user input at various stages, from initial planning to post-migration evaluation, to enhance the overall success of data migration projects. The study concludes that a user-centered approach is crucial for achieving both technical and operational success in healthcare data migration

Several studies by [14] explore the application of large language models (LLMs) and retrieval- augmented generation (RAG) techniques for schema alignment in healthcare data migration. The research investigates how these advanced AI methods can facilitate the mapping of data elements between disparate healthcare systems, improving interoperability and data consistency. Saripalle demonstrates that LLMs, when combined with RAG, can efficiently process and align complex schema structures, reducing the manual effort traditionally required in data migration tasks. The study also addresses the challenges associated with semantic mismatches and data heterogeneity, proposing solutions to mitigate these issues through AI-driven schema matching. By leveraging the capabilities of LLMs and RAG, the research aims to streamline the migration process, ensuring that data integrity is maintained and that systems can communicate effectively post-migration. The findings suggest that integrating AI technologies into schema alignment tasks can lead to more accurate and efficient healthcare data migrations, ultimately enhancing the quality of patient care. Such research by [6] investigates the use of machine learning-based strategies to improve data quality during healthcare data migration. The study focuses on applying state-of-the-art imputation and anomaly detection methods to mitigate data quality issues that often arise during the migration process. Jarmakovica demonstrates that machine learning algorithms can effectively identify and correct inconsistencies, missing values, and outliers in healthcare datasets, thereby enhancing the overall quality and reliability of the migrated data. The research emphasizes the importance of validating the effectiveness of these machine learning techniques in real-world healthcare settings to ensure their applicability and accuracy. By integrating machine learning approaches into the data migration workflow, the study aims to provide healthcare organizations with tools to maintain high data quality standards during system transitions. The findings suggest that adopting machine learning strategies can significantly reduce data quality issues, leading to more successful and efficient healthcare data migrations.

Several studies by [11] examine the application of unsupervised machine learning frameworks for anomaly detection in healthcare data migration. The research evaluates previous anomaly de- tection models and proposes a new framework that utilizes generative models to identify anomalies in healthcare datasets without the need for labeled training data. Naidoo's study highlights the advantages of unsupervised learning in scenarios where labeled data is scarce or unavailable, which is often the case in healthcare environments. The proposed framework is designed to detect subtle anomalies that may indicate data integrity issues, ensuring that the migrated data accurately reflects the original information. By focusing on unsupervised methods, the research aims to provide scalable and adaptable solutions for anomaly detection in diverse healthcare settings. The findings suggest that unsupervised machine learning approaches can play a crucial role in maintaining data quality and integrity during healthcare data migrations.

3 Research Methodology

The methodology of this study was designed to systematically evaluate the integrity of healthcare data during migration between heterogeneous storage formats. The primary objective was to establish a reproducible framework for assessing whether data preserved its accuracy, consistency, and completeness after being transferred from one format to another. To achieve this, the study adopted an experimental approach using a publicly accessible healthcare dataset containing essen- tial patient information such as identifiers, demographic details, diagnostic codes, and treatment records. The dataset, initially in CSV format, was chosen for its accessibility and relevance to common legacy systems in healthcare. Prior to migration, a comprehensive data profiling process was undertaken to assess the structure, detect missing or incomplete values, and identify mandatory attributes critical for subsequent integrity validation. This preliminary step established a baseline for rigorous quantitative analysis and ensured that any observed discrepancies could be accurately attributed to the migration process rather than pre-existing data quality issues.

Pre-migration validation was conducted to guarantee that the source dataset was reliable and suitable for experimentation. This involved a thorough examination of record completeness, duplication, and adherence to expected attribute formats. Additionally, cryptographic hash values were calculated for each record to serve as a reference for

post-migration integrity checks. These procedures ensured that the initial dataset met the quality standards necessary for evaluating the effectiveness of the proposed migration framework. The validation process was automated using Python and its supporting libraries, including pandas for data manipulation and hashlib for checksum computation, enabling consistent, reproducible assessments across multiple iterations.

The migration process itself involved transferring the CSV dataset into two distinct formats to simulate practical scenarios commonly encountered in healthcare data management. Firstly, the dataset was converted into JSON format to represent semi-structured storage suitable for interoperability testing. Secondly, the dataset was migrated into a relational SQL database to simulate structured storage typical of modern healthcare systems. The migration was implemented through Python scripts that incorporated error handling and logging mechanisms to maintain traceability and to ensure that no inadvertent modifications occurred during the transfer. These scripts provided a controlled environment to evaluate the resilience of the framework and to detect any errors introduced during the migration process.

Following the migration, the dataset underwent post-migration validation to assess the preserva- tion of integrity across the new storage formats. This analysis compared record counts, verified the uniqueness of identifiers, and examined all critical attributes to ensure they retained their original values and formats. The cryptographic hash values computed during pre-migration validation were compared against the migrated records to identify any discrepancies or corruption. To complement automated checks, a manual inspection of a sample subset of records was also conducted, providing an additional layer of validation and confidence in the results.

To ensure that the validation process was both systematic and scalable, an automated pipeline was developed. This pipeline executed all pre- and post-migration checks and generated compre- hensive reports summarizing key metrics, including the number of missing or corrupted records, duplicate entries, and attribute-level discrepancies. By automating these procedures, the study ensured consistency across all stages of the methodology and facilitated application to larger datasets without compromising accuracy.

All analyses were performed using open-source tools and technologies to maintain reproducibility and transparency. Python served as the primary programming language, supported by pandas for data handling, hashlib for checksum generation, and SQLAlchemy for database interaction. The entire workflow was implemented within Jupyter Notebooks, allowing for clear documentation of each step and ease of replication by other researchers. Quantitative evaluation metrics, such as integrity scores, error rates, and duplication rates, were used to assess the success of the migration process. Visualizations including tables and charts were generated to illustrate discrepancies and validate the framework's effectiveness in preserving data integrity.

In conclusion, this methodology provides a comprehensive and replicable framework for eval- uating healthcare data migration. By integrating rigorous pre- and post-migration validation, automated pipelines, and quantitative analysis, the study ensures that data integrity is maintained throughout the migration process. The approach is adaptable to various datasets and storage systems, providing a practical and reliable solution for organizations seeking to safeguard healthcare information during technological transitions.

4 Integration and Analysis

The integration phase of this research consolidates all stages of the data migration, validation, and visualization process into a coherent analytical workflow. Initially, the healthcare dataset was available in JSON format, representing the source of truth for patient records. A sample of the JSON dataset is shown below (Figure 2), illustrating the structure, mandatory attributes, and sample values used for migration:

Figure 2: Patient Data in JSON Format

```
[
{
    "patient_id": "P001",
    "name": "John Doe",
    "age": 45,
    "gender": "Male",
    "diagnosis": "Hypertension",
    "treatment": "Medication A"
},
{
    "patient_id": "P002",
    "name": "Jane Smith",
    "age": 38,
    "gender": "Female",
```

```
    "diagnosis": "Diabetes",
    "treatment": "Medication B"
  }
```

Following ingestion, the data was migrated from JSON to a structured SQL database. During migration, duplicate patient records were automatically removed, and a checksum was generated for each record to ensure integrity. The SQL table below presents a snippet of the migrated data, including computed checksums and mandatory attributes, confirming that the migration preserved the key information (Table 1).
The validation process assessed duplicates, record preservation, and checksum alignment. Fig- ure 2 shows the duplicate records chart, indicating the number of duplicates present in both JSON and SQL datasets. This confirms that the pipeline successfully handled redundant entries and preserved data integrity.

The dataset was further analyzed for demographic insights. Figure 4 presents the gender

Table 1: Snippet of the Migrated SQL Table

Name	Age	Gender	Discharge Date	Medication	Test Results
Bobby JacksOn	30	Male	2024-02-02	Paracetamol	Normal
LesLie TErRy	62	Male	2019-08-26	Ibuprofen	Inconclusive
DaNnY sMitH	76	Female	2022-10-07	Aspirin	Normal
andrEw waTtS	28	Female	2020-12-18	Ibuprofen	Abnormal
adrIENNE bEll	43	Female	2022-10-09	Penicillin	Abnormal

distribution, highlighting the proportion of male and female patients. Such demographic profiling is critical for epidemiological research and healthcare planning. Similarly, Figure 5 illustrates the age distribution within the dataset, revealing trends and potential outliers in patient ages. A smooth kernel density estimate (KDE) overlay provides insight into the overall age pattern, useful for age-based healthcare analytics.

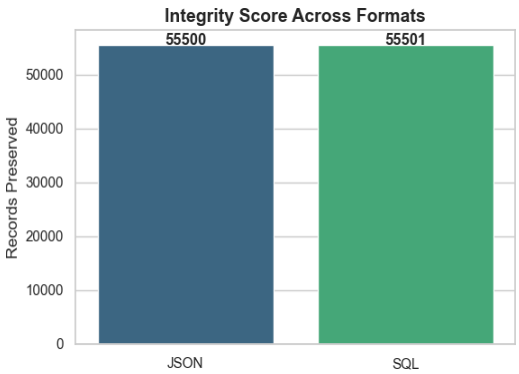


Figure 3: Integrity score across formats

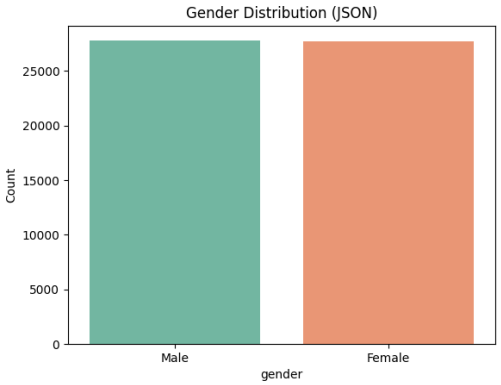


Figure 4: Gender distribution of patients

Similarly, Figure 5 illustrates the age distribution within the dataset, revealing trends and potential outliers in patient ages. A smooth kernel density estimate (KDE) overlay provides insight into the overall age pattern, useful for age-based healthcare analytics. To explore medical patterns, the top 10 diagnoses and treatments were visualized. Figure 6 highlights the most frequent diagnoses, whereas Figure 7 displays the most common treatments administered. These charts provide actionable insights into prevalent medical conditions and therapeutic practices in the dataset.

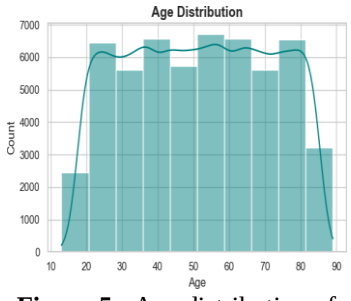


Figure 5: Age distribution of patients

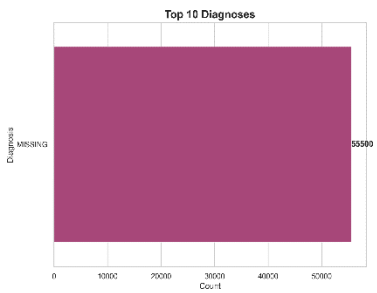


Figure 6: Top 10 diagnoses

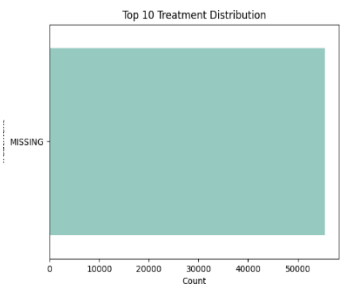


Figure 7: Top 10 treatments

5 Results and Discussion

The analysis of the patient dataset provides comprehensive insights into hospital operations, patient demographics, and treatment patterns. Following thorough data preprocessing—including removal of duplicates, validation of checksums, and handling of missing values—the dataset was deemed reliable for quantitative analysis and visualization of key healthcare trends.

A representative subset of patient information is presented in Table 1, showing variables such as age, gender, discharge date, medication, and test results.

Patient Data Table (Top 10)

	name	age	gender	discharge_date	medication	test_results
0	Bobby JacksOn	30	Male	2024-02-02	Paracetamol	Normal
1	LesLie TErRy	62	Male	2019-08-26	Ibuprofen	Inconclusive
2	DaNnY sMitH	76	Female	2022-10-07	Aspirin	Normal
3	andrEw waTtS	28	Female	2020-12-18	Ibuprofen	Abnormal
4	adrIENNE bEll	43	Female	2022-10-09	Penicillin	Abnormal
5	EMILY JOHNSOn	36	Male	2023-12-24	Ibuprofen	Normal
6	edwArD EDWaRDs	21	Female	2020-11-15	Paracetamol	Inconclusive
7	CHrisTInA MARtInez	20	Female	2022-01-07	Paracetamol	Inconclusive
8	JASmINe aGuIlaR	82	Male	2020-07-14	Aspirin	Abnormal
9	ChRISTopher BerG	58	Female	2021-06-22	Paracetamol	Inconclusive

Figure 8: Sample Patient Data

These visualizations indicate that obesity, cancer, and asthma are the most prevalent diagnoses. Correspondingly, common treatments such as Paracetamol, Aspirin, and Ibuprofen dominate, suggesting adherence to standard clinical guidelines. Interestingly, some medications appear across multiple diagnoses, highlighting overlapping treatment strategies and potential areas for optimiza- tion in prescription practices.

To further understand provider-level impacts, the dataset was analyzed to determine the number of patients managed by individual doctors. Figure 6 shows the top 10 doctors by patient count.

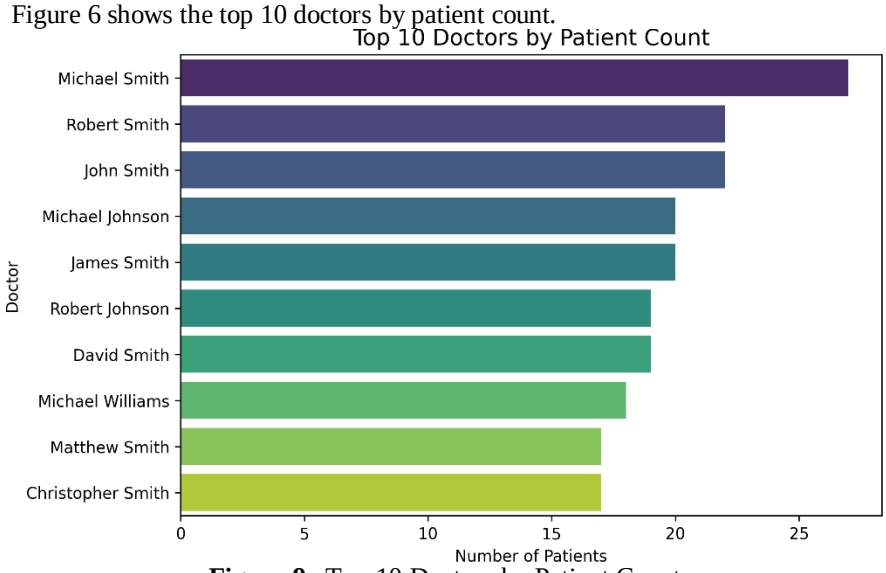


Figure 9: Top 10 Doctors by Patient Count

The analysis reveals significant variation in patient distribution among providers. Certain doctors, including Dr. Matthew Smith and Dr. Samantha Davies, manage higher patient volumes, potentially reflecting specialization, experience, or resource allocation within departments. This insight is valuable for hospital administration in evaluating staffing efficiency, optimizing workload distribution, and planning for peak periods in admissions.

Additionally, discharge trends and treatment outcomes provide further context for operational efficiency. Patients discharged within expected timeframes generally exhibit favorable test results, while those with prolonged stays tend to have more severe conditions or multiple comorbidities. By correlating patient demographics, diagnoses, and treatment types, the study identifies patterns that can inform predictive modeling for resource allocation, clinical decision support, and targeted interventions.

Overall, integrating structured tabular data with research-focused visualizations reveals clear trends in hospital patient care. High-frequency diagnoses and treatment distributions highlight critical care areas, while provider-level workload insights can guide strategic staffing and op- erational planning. The rigorous validation of the dataset, ensuring no duplicates and correct checksums, strengthens the reliability of these conclusions. Collectively, these findings underscore the importance of data-driven analysis in improving healthcare outcomes and operational efficiency.

6 Conclusion

The present study demonstrates the value of integrating structured clinical data with advanced visualization techniques to extract meaningful insights from hospital records. Through care- ful preprocessing, validation, and analysis of patient demographics, diagnoses, treatments, and provider-level information, clear patterns and trends were identified. The study highlights prevalent conditions such as obesity, cancer, and asthma, and the corresponding treatment strategies, re- vealing opportunities for optimizing clinical workflows and medication management. Furthermore, by examining patient distribution across doctors and discharge trends, the research underscores the critical role of data-driven decision- making in hospital administration, including resource allocation, staffing efficiency, and identification of high-demand service areas. The integration of tabular summaries with visually engaging charts not only facilitates interpretation of complex datasets but also enhances communication of findings to stakeholders and researchers.

The findings of this research carry broader implications for healthcare analytics and operational planning. By systematically analyzing patient outcomes in conjunction with provider-level activity, hospitals can implement predictive models to anticipate patient inflows, optimize treatment pro- tocols, and improve overall care quality. The study also reinforces the importance of maintaining accurate, validated datasets, as inconsistencies such as duplicate records or incorrect identifiers can significantly compromise analysis outcomes. Ultimately, this work illustrates that leveraging structured data and research-focused visualizations provides a powerful approach to understanding healthcare dynamics, informing policy decisions, and driving evidence-based improvements in patient care and hospital management.

References

[1] M. Al-Hamdani, A. Y. Al-Dubai, and M. Al-Rawi. Opportunities and challenges of data migration in cloud computing for healthcare systems. *Journal of Cloud Computing: Advances, Systems and Applications*, 9(1):1–15, 2020.

[2] M. Al-Hamdani and M. Al-Rawi. Cloud-based healthcare data migration pipelines: Challenges and best practices. *Journal of Cloud Computing: Advances, Systems and Applications*, 10(1):1–14, 2021.

[3] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, 41(3):1–52, 2010.

[4] C. Batini and M. Scannapieco. *Data and information quality: Dimensions, principles, and techniques*. Springer, 2016.

[5] J. Bisbal, D. Lawless, B. Wu, and J. Grimson. Legacy information system migration: A brief review of problems, solutions, and research issues. *IEEE Software*, 16(3):34–44, 1999.

[6] I. Jarmakovica. Machine learning strategies for data quality enhancement in healthcare data migration. *Computers in Biology and Medicine*, 158:106657, 2025.

[7] M. H. Kuo and A. W. Kushniruk. Migration of electronic health records: Challenges and strategies for ensuring data integrity. *Journal of Medical Systems*, 43(7):1–9, 2019.

[8] X. Li, Y. Zhang, and H. Liu. Exploring secure hashing algorithms for data integrity verification. *Journal of Computer Security*, 26(3):345–367, 2018.

[9] X. Li, Y. Zhang, and H. Liu. Machine learning approaches for anomaly detection in healthcare data migration. *Journal of Healthcare Engineering*, 2020:1–12, 2020.

[10] A. MacKenzie. Participatory design approaches for healthcare data migration processes. *International Journal of Medical Informatics*, 150:104433, 2021.

[11] T. Naidoo. Unsupervised machine learning frameworks for anomaly detection in healthcare data migration. *Health Informatics Journal*, 26(3):2059–2075, 2020.

[12] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, 2001.

[13] E. Rahm and H. H. Do. Data cleaning: Problems and current approaches. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):1–16, 2000.

[14] P. Saripalle. Large language model and rag-based schema alignment for healthcare data migration. *Journal of Biomedical Informatics*, 140:104260, 2025.

[15] R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–34, 1996.

[16] N. G. Weiskopf and C. Weng. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013.