Healthcare Analytics System

Dataset Documentation Report

A Comprehensive Guide to Data Sources, Structure, and Preprocessing

Project	Healthcare Predictive Analytics System
Total Datasets	5 CSV files
Total Records	6,621 records
Time Period	January 2022 - December 2024
Primary Purpose	Patient Readmission Risk Prediction

1. Executive Summary

This documentation provides a comprehensive overview of the healthcare analytics dataset used for developing a machine learning-based patient readmission risk prediction system. The dataset ecosystem consists of five interconnected CSV files that collectively capture patient demographics, physician performance metrics, departmental operations, financial indicators, and physician registry information. These datasets span a three-year period from January 2022 through December 2024, providing substantial historical data for trend analysis and predictive modeling.

The primary objective of this data collection effort was to enable healthcare administrators and clinical staff to identify patients at high risk of hospital readmission within 30 days of discharge. By combining demographic factors, clinical outcomes, and operational metrics, the system can generate actionable insights that support proactive patient care interventions and resource allocation decisions. The datasets have been carefully structured to maintain referential integrity while supporting both descriptive analytics and machine learning applications.

All five datasets underwent rigorous quality assurance processes during creation, including validation of data types, range checking for numerical fields, and verification of categorical consistency. The resulting data infrastructure supports the healthcare system's goal of reducing preventable readmissions while maintaining high standards of patient care and operational efficiency.

2. Dataset Overview

The healthcare analytics system relies on five primary datasets, each serving a distinct purpose within the analytical framework. The following table summarizes the structure and scope of each dataset, providing a foundation for the detailed explanations that follow in subsequent sections.

Dataset File	Records	Columns	Primary Purpose
patient_demographics.csv	1,001	7	Patient characteristics and outcomes
physician_performance.csv	3,960	10	Monthly physician metrics
department_metrics.csv	612	10	Department operational data
financial_performance.csv	36	10	Hospital-wide financials
physician_registry.csv	110	7	Physician master directory

The datasets are interconnected through common identifiers, most notably the physician_id and department_id fields, which enable cross-referential analysis. Time-based fields (month and year) provide temporal alignment across datasets, supporting longitudinal studies of trends and patterns. This relational structure was deliberately designed to support complex analytical queries while maintaining data normalization principles that minimize redundancy and ensure consistency.

3. Patient Demographics Dataset

File: patient_demographics.csv | Records: 1,001 | Columns: 7

The patient demographics dataset serves as the primary data source for machine learning model training and represents aggregated patient information grouped by demographic characteristics. This dataset was created through a systematic extraction process from the hospital's electronic health records system, with patient-level data aggregated to protect individual privacy while preserving analytical utility. Each record represents a unique combination of age group, gender, and insurance type, with associated outcome metrics calculated as averages across all patients within that demographic segment.

The creation process involved several key steps. First, raw patient encounter data was extracted from the hospital information system, including admission dates, discharge dates, diagnosis codes, and billing information. Second, patients were categorized into five age groups (0-17, 18-29, 30-49, 50-64, and 65+) based on their age at the time of admission. Third, demographic groupings were formed by combining age group, gender (Male or Female), and insurance provider. Finally, outcome metrics were calculated for each group, including average length of stay, average treatment cost, and the proportion of patients who were readmitted within 30 days of discharge.

Data Dictionary:

Column Name	Data Type	Description
age_group	Categorical	Patient age range (0-17, 18-29, 30-49, 50-64, 65+)
gender	Categorical	Patient gender (M = Male, F = Female)
insurance_type	Categorical	Insurance provider name (10 providers)
patient_count	Integer	Number of patients in this demographic group
avg_length_of_stay	Float	Mean hospital stay duration in days
avg_cost	Float	Mean treatment cost in US dollars
readmission_rate	Float	Proportion readmitted within 30 days (0.0-1.0)

The preprocessing pipeline for this dataset involved several transformations to prepare the data for machine learning applications. Categorical variables (age_group, gender, and insurance_type) were encoded using scikit-learn's LabelEncoder, which assigns integer values to each unique category. This encoding preserves the categorical nature of the data while enabling numerical computation required by machine learning algorithms. The age_group field was encoded as integers 0 through 4, gender as 0 or 1, and insurance_type as integers 0 through 9 corresponding to the ten insurance providers.

The target variable for classification was created by applying a threshold to the readmission_rate field. Patients with a readmission rate exceeding 0.20 (20 percent) were classified as high risk (labeled as 1), while those at or below this threshold were classified as low risk (labeled as 0). This threshold was selected based on clinical guidance indicating that a 20 percent readmission rate represents a meaningful boundary for intervention prioritization. The resulting class distribution showed 33 percent of records classified as high risk and 67 percent as low risk, creating a moderate class imbalance that required additional handling through SMOTE (Synthetic Minority Over-sampling Technique) during model training.

The ten insurance providers represented in this dataset include Aetna, Anthem, Blue Cross Blue Shield, Cigna, Health Net, Humana, Kaiser Permanente, Medicaid, Medicare, Molina Healthcare, Tricare, and UnitedHealthcare. This diverse mix of private insurers, government programs (Medicare and Medicaid), and managed care organizations reflects the typical payer mix encountered in a large healthcare system. Insurance type was included as a feature because research has shown correlations between insurance coverage and healthcare utilization patterns, which may influence readmission risk.

4. Physician Performance Dataset

File: physician_performance.csv | Records: 3,960 | Columns: 10

The physician performance dataset captures monthly performance metrics for 110 physicians across all hospital departments. This dataset was created to support quality improvement initiatives and enable performance benchmarking at the individual provider level. Data collection occurs on a monthly basis, with metrics calculated from electronic health records, billing systems, and patient satisfaction surveys. The 36-month time span (January 2022 through December 2024) multiplied by 110 physicians yields the 3,960 records contained in this dataset.

The creation process for this dataset involves automated extraction from multiple source systems at the end of each calendar month. Patient encounter data is aggregated by attending physician to calculate volume metrics (total patients) and clinical outcomes (length of stay, complication rate, readmission rate). Revenue data is extracted from the billing system and averaged across each physician's patient population. Patient satisfaction scores are compiled from post-discharge surveys, with responses mapped to a standardized 1-to-5 scale where 1 represents very dissatisfied and 5 represents very satisfied.

Data Dictionary:

Column Name	Data Type	Description
physician_id	String	Unique identifier (PHY000 through PHY109)
physician_name	String	Full name with professional title (Dr.)
month	Integer	Calendar month (1-12)
year	Integer	Calendar year (2022, 2023, or 2024)
total_patients	Integer	Number of patients treated during the month
avg_length_of_stay	Float	Mean patient stay duration in days
avg_satisfaction_score	Float	Mean patient satisfaction (1.0-5.0 scale)
complication_rate	Float	Proportion of cases with complications (0.0-1.0)
readmission_rate	Float	Proportion readmitted within 30 days (0.0-1.0)
avg_revenue	Float	Mean revenue per patient in US dollars

Preprocessing for this dataset focuses on temporal alignment and normalization. The month and year fields enable time-series analysis and trend detection, allowing administrators to track physician performance trajectories over time. Rate fields (complication_rate and readmission_rate) are stored as decimal proportions rather than percentages to maintain consistency with statistical conventions and facilitate direct use in calculations. The satisfaction score uses a bounded 1-to-5 scale that has been validated through psychometric testing to ensure reliable measurement of patient experience.

Quality assurance processes applied to this dataset include range validation for all numerical fields, verification that physician_id values match entries in the physician registry, and consistency checks to ensure that monthly records exist for all active physicians. Missing data is rare in this dataset because the automated extraction process requires complete information for each field. In cases where a physician had no patient encounters during a month (such as during extended leave), no record is generated rather than creating a record with zero values, which could distort aggregate statistics.

5. Physician Registry Dataset

File: physician_registry.csv | Records: 110 | Columns: 7

The physician registry dataset serves as the master directory for all physicians affiliated with the healthcare system. This reference dataset was created to maintain consistent physician identification across all operational and analytical systems while providing essential demographic and organizational information. Unlike the time-series datasets, the physician registry represents a point-in-time snapshot of the physician workforce, updated whenever staffing changes occur such as new hires, departures, or departmental transfers.

The dataset creation process involves integration with the human resources information system and medical staff credentialing database. Each physician receives a unique identifier (physician_id) upon joining the medical staff, which persists throughout their tenure and serves as the primary key for linking to performance data. Name fields are separated into first_name and last_name components to support flexible display formatting while the physician_name field provides the formatted full name with professional title for reporting purposes.

Data Dictionary:

Column Name	Data Type	Description
physician_id	String	Unique identifier (PHY000 through PHY109)
physician_name	String	Full name with title (e.g., Dr. Sarah Johnson)
first_name	String	Physician first name
last_name	String	Physician last name
specialty	String	Medical specialty (17 specialties represented)
department_id	String	Department identifier (DEPT000 through DEPT016)
hospital	String	Primary hospital affiliation

The specialty field categorizes physicians into one of 17 medical specialties: Emergency Medicine, Cardiology, Orthopedics, Neurology, Oncology, Pediatrics, Internal Medicine, Surgery, Psychiatry, Radiology, Anesthesiology, Obstetrics, Gastroenterology, Dermatology, Urology, Nephrology, and Pulmonology. Each specialty aligns with a corresponding department, establishing the organizational hierarchy that connects individual physicians to departmental performance metrics. The department_id field serves as a foreign key linking to the department_metrics dataset.

The hospital field indicates the primary facility where each physician practices, with eight hospitals represented in the system: City General Hospital, St. Mary's Medical Center, University Health System, Memorial Hospital, Regional Medical Center, Community Health Hospital, Metropolitan Medical Center, and Riverside Hospital. This multi-facility structure reflects a regional healthcare network where physicians may have privileges at multiple locations but maintain primary affiliation with one facility for administrative and scheduling purposes.

Preprocessing for the physician registry focuses on standardization and validation rather than transformation. Name fields are validated against standard naming conventions, and specialty assignments are verified against the approved specialty list to prevent data entry errors. The registry serves as the authoritative source for physician information, with changes propagated to dependent systems through automated synchronization processes that maintain referential integrity across the data ecosystem.

6. Department Metrics Dataset

File: department_metrics.csv | Records: 612 | Columns: 10

The department metrics dataset provides monthly operational statistics for the 17 clinical departments within the healthcare system. This dataset was created to support capacity planning, resource allocation, and departmental performance evaluation. Data is aggregated from individual patient encounters, staffing records, and financial systems to provide a comprehensive view of departmental operations. The 36-month time span multiplied by 17 departments yields 612 records, with each record representing one department's performance during a specific month.

The creation process involves end-of-month batch processing that aggregates encounter-level data to the department level. Admission counts are tallied from the patient registration system, while length of stay and cost metrics are calculated as averages across all patients discharged during the month. Revenue figures represent actual collections plus accrued receivables attributed to each department. Occupancy rates are derived from daily census data, calculating the proportion of available beds occupied on average throughout the month. Nurse-to-patient ratios are calculated from staffing schedules and patient census records.

Data Dictionary:

Column Name	Data Type	Description
department_id	String	Unique identifier (DEPT000 through DEPT016)
department_name	String	Full department name
month	Integer	Calendar month (1-12)
year	Integer	Calendar year (2022, 2023, or 2024)
total_admissions	Integer	Number of patient admissions during the month
avg_length_of_stay	Float	Mean patient stay duration in days
avg_cost	Float	Mean cost per admission in US dollars
total_revenue	Float	Total department revenue for the month
occupancy_rate	Float	Average bed occupancy proportion (0.0-1.0)
nurse_patient_ratio	Float	Average nurses per patient

The 17 departments span the full range of clinical services offered by the healthcare system. Emergency Medicine (DEPT000) serves as the primary entry point for acute care patients, while specialty departments such as Cardiology (DEPT001), Orthopedics (DEPT002), and Neurology (DEPT003) provide focused expertise for specific conditions. Support departments including Radiology (DEPT009) and Anesthesiology (DEPT010) provide essential services that span multiple clinical pathways. This comprehensive departmental structure enables detailed analysis of resource utilization patterns and identification of optimization opportunities.

Preprocessing considerations for this dataset include validation of occupancy rates to ensure values remain within the valid 0.0 to 1.0 range, verification that nurse-to-patient ratios meet minimum staffing standards, and reconciliation of revenue figures against financial system totals. Time-series preprocessing enables trend analysis by ordering records chronologically and calculating month-over-month and year-over-year changes for key metrics. Seasonal adjustment factors can be applied to account for predictable variations in patient volumes throughout the calendar year.

7. Financial Performance Dataset

File: financial_performance.csv | Records: 36 | Columns: 10

The financial performance dataset provides monthly hospital-wide financial metrics aggregated across all departments and service lines. This dataset was created to support executive-level financial monitoring, budget variance analysis, and strategic planning activities. Unlike the department-level financial data, this dataset presents a consolidated view of organizational financial health, enabling assessment of overall profitability, liquidity, and operational efficiency.

The creation process involves monthly closing procedures where departmental financials are rolled up to the organizational level. Revenue figures include patient service revenue, ancillary revenue, and other operating revenue. Expense categories are summarized into total expenses, which encompass salaries, supplies, purchased services, and overhead costs. Deductions from revenue are tracked separately, including bad debt (uncollectable patient balances), charity care (services provided without charge to qualifying patients), and insurance contractual adjustments (the difference between charges and contracted payment rates).

Data Dictionary:

Column Name	Data Type	Description
month	Integer	Calendar month (1-12)
year	Integer	Calendar year (2022, 2023, or 2024)
total_revenue	Float	Total operating revenue in US dollars
total_expenses	Float	Total operating expenses in US dollars
net_income	Float	Revenue minus expenses (profit/loss)
operating_margin	Float	Net income divided by revenue (profitability ratio)
bad_debt	Float	Uncollectable patient balances written off
charity_care	Float	Value of uncompensated care provided
insurance_contractual	Float	Insurance rate adjustments
cash_on_hand	Float	Available liquid assets

Over the three-year period captured in this dataset, the healthcare system generated approximately 709 million dollars in total revenue while incurring 575 million dollars in expenses, resulting in net income of approximately 134 million dollars. The operating margin averaged 18.8 percent across the period, though individual months ranged from 7.6 percent to 27.8 percent depending on patient volumes, payer mix, and expense timing. Cash on hand averaged 5.8 million dollars, providing adequate liquidity for operational needs while maintaining prudent reserve levels.

Preprocessing for financial data focuses on validation and normalization. All monetary values are stored in US dollars without currency symbols to facilitate numerical operations. The operating margin is pre-calculated and stored to ensure consistent definition across analyses (net income divided by total revenue). Range validation ensures that operating margin values fall within reasonable bounds (typically -50 percent to +50 percent for healthcare organizations) and that cash on hand remains positive. Year-over-year comparison calculations are facilitated by the standardized month-year structure shared across all time-series datasets.

8. Data Preprocessing Pipeline

The data preprocessing pipeline transforms raw healthcare data into analysis-ready formats suitable for both descriptive analytics and machine learning applications. This pipeline was developed using Python with the pandas library for data manipulation and scikit-learn for machine learning preparation. The pipeline operates in sequential stages, each building upon the outputs of previous stages to produce the final analytical datasets.

Stage 1: Data Loading and Validation

The first stage loads CSV files into pandas DataFrames and performs comprehensive validation checks. Column names are verified against expected schemas, data types are validated (ensuring numerical fields contain only numeric values), and missing value analysis is performed. For this dataset collection, validation confirmed zero missing values across all five files, eliminating the need for imputation strategies. Range validation ensures that rate fields (readmission_rate, occupancy_rate, complication_rate) fall within the valid 0.0 to 1.0 range and that date fields contain valid month (1-12) and year (2022-2024) values.

Stage 2: Target Variable Creation

For machine learning applications, a binary target variable is created from the continuous readmission_rate field in the patient demographics dataset. The transformation applies a threshold of 0.20, classifying records with readmission rates above 20 percent as high risk (value 1) and those at or below 20 percent as low risk (value 0). This threshold was determined through clinical consultation and represents a meaningful boundary for intervention prioritization. The resulting class distribution showed 334 high-risk records (33.4 percent) and 667 low-risk records (66.6 percent), creating a moderate class imbalance that required additional handling in subsequent stages.

Stage 3: Categorical Encoding

Categorical variables require numerical encoding for use in machine learning algorithms. The pipeline employs scikit-learn's LabelEncoder to transform categorical fields into integer representations. The age_group field is encoded as integers 0 through 4 corresponding to the five age categories. The gender field is encoded as 0 (Female) or 1 (Male). The insurance_type field is encoded as integers 0 through 9 representing the ten insurance providers. Each encoder is persisted to enable inverse transformation when interpreting model outputs or applying the pipeline to new data.

Stage 4: Train-Test Split and SMOTE Balancing

The preprocessed data is split into training and testing subsets using a 75-25 ratio with stratified sampling to preserve class proportions. The training set initially contained 751 samples with the same 67-33 class imbalance as the full dataset. To address this imbalance, Synthetic Minority Over-sampling Technique (SMOTE) is applied to the training data only. SMOTE generates synthetic examples of the minority class by interpolating between existing minority class samples, using k=5 nearest neighbors for interpolation. After SMOTE application, the training set expands to 1,006 samples with a perfectly balanced 50-50 class distribution (503 samples per class). The test set remains unmodified to provide an unbiased evaluation of model performance on naturally distributed data.

9. Feature Engineering

Feature engineering transforms the preprocessed variables into the final feature set used for model training. The patient demographics dataset provides six features for the readmission risk prediction model: three encoded categorical features (age_group_encoded, gender_encoded, insurance_type_encoded) and three numerical features (patient_count, avg_length_of_stay, avg_cost). The readmission_rate field is excluded from features as it serves as the basis for the target variable.

Final Feature Set:

Feature	Type	Encoding	Range
age_group_encoded	Categorical	Label (0-4)	0, 1, 2, 3, 4
gender_encoded	Categorical	Label (0-1)	0, 1
insurance_type_encoded	Categorical	Label (0-9)	0 through 9
patient_count	Numerical	None (raw)	12 to 284
avg_length_of_stay	Numerical	None (raw)	2.0 to 7.9 days
avg_cost	Numerical	None (raw)	$3,017 to $12,977

Feature importance analysis conducted after model training revealed that avg_cost is the most predictive feature for readmission risk, contributing approximately 26 percent of model predictive power in the Random Forest model. This finding aligns with clinical intuition, as higher costs often correlate with more complex cases that carry elevated readmission risk. Patient_count ranks second at 22 percent, suggesting that the size of demographic segments provides meaningful signal about population-level risk patterns. Average length of stay contributes 19 percent, reflecting the established relationship between extended hospitalizations and subsequent readmission likelihood.

The categorical features collectively contribute the remaining 33 percent of predictive power, with insurance type providing 14 percent, age group providing 11 percent, and gender providing 8 percent. The relatively strong contribution of insurance type may reflect differences in care patterns, coverage limitations, or socioeconomic factors associated with different payer populations. Age group's contribution aligns with known relationships between age and health complexity, while gender's smaller contribution suggests limited sex-based differences in readmission risk within this patient population.

10. Conclusions

This documentation has provided a comprehensive overview of the five datasets comprising the healthcare analytics system's data infrastructure. The patient demographics dataset serves as the primary source for machine learning model training, with 1,001 records capturing aggregated patient characteristics and outcomes across demographic segments. The physician performance dataset enables individual provider analysis through 3,960 monthly performance records spanning 110 physicians. The department metrics dataset supports operational analysis with 612 monthly records across 17 departments. The financial performance dataset provides executive-level visibility into organizational financial health through 36 monthly records. The physician registry serves as the authoritative directory linking physicians to their specialties, departments, and hospital affiliations.

The preprocessing pipeline transforms these raw datasets into analysis-ready formats through systematic validation, encoding, and balancing procedures. Key preprocessing steps include categorical encoding using LabelEncoder, binary target creation using a clinically-informed 20 percent threshold, stratified train-test splitting, and SMOTE-based class balancing to address the natural imbalance in readmission outcomes. The resulting feature set of six variables provides meaningful predictive signal while maintaining interpretability for clinical stakeholders.

The data infrastructure supports the healthcare system's strategic goal of reducing preventable readmissions through proactive identification of high-risk patients. By combining demographic characteristics with clinical and operational metrics, the system enables targeted interventions that can improve patient outcomes while optimizing resource utilization. The relational structure connecting physicians, departments, and financial metrics provides additional context for root cause analysis and performance improvement initiatives.

Future enhancements to this data infrastructure may include integration of additional clinical features such as diagnosis codes, procedure codes, and medication histories; expansion of temporal granularity to enable daily or weekly trend analysis; and incorporation of social determinants of health data to improve prediction accuracy for vulnerable populations. The modular design of the preprocessing pipeline facilitates these extensions while maintaining backward compatibility with existing analytical applications.

— End of Documentation —