Synthetic EDI-278 Prior Authorization Dataset

Technical Demonstration of Data Engineering Capabilities

# Executive Summary

This document outlines the comprehensive methodology used to generate a fully synthetic EDI-278 prior authorization dataset for demonstrating technical expertise in healthcare data engineering. The dataset contains 600 records spanning all 50 US states where Centene Corporation operates, with realistic distributions and business logic that mirrors actual healthcare prior authorization processes.

The primary objective of this project is to showcase advanced data engineering capabilities including synthetic data generation, statistical modeling, data visualization, and dashboard development - all while maintaining complete confidentiality and data safety.

# Data Generation Methodology

## 1. Dataset Architecture

The synthetic dataset is built on a comprehensive schema that captures all critical aspects of EDI-278 prior authorization transactions:

### Identifiers

* request\_id: Unique request identifier (REQ + 8 alphanumeric)
* prior\_auth\_id: Prior authorization ID (PA + 8 alphanumeric)
* submitter\_id: Submitting entity ID (SUB + 8 alphanumeric)
* provider\_npi\_masked: Masked provider NPI (\*\*\*\*\*\*\*\*1234 format)
* facility\_id: Facility identifier (FAC + 8 alphanumeric)
* member\_id\_masked: Masked member ID (MBR + 6 digits)

### Temporal Fields

* request\_timestamp: When the prior auth request was submitted (ISO format)
* decision\_timestamp: When the decision was made (ISO format)
* turnaround\_hours: Hours between request and decision (integer)

### Status & SLA

* sla\_met: Whether SLA was met (1=yes, 0=no)
* status: Final status (received, pended, approved, denied)
* urgent\_flag: Whether request was marked urgent (1=yes, 0=no)

### Request Details

* request\_type: Type of request (initial, concurrent, extension)
* service\_type: Type of service requested
* place\_of\_service: Where service will be provided
* state\_code: US state code (2-letter)
* state\_name: Full state name
* health\_plan: Centene health plan name for the state
* line\_of\_business: Medicaid, Medicare, or Marketplace
* age\_band: Patient age group

### Resolution Details

* pend\_reason: Reason for pended status (if applicable)
* deny\_reason: Reason for denial (if applicable)
* resubmission\_flag: Whether this is a resubmission (1=yes, 0=no)
* appeal\_flag: Whether an appeal was filed (1=yes, 0=no)

## 2. Statistical Modeling Approach

The data generation employs sophisticated statistical modeling to ensure realistic distributions and business logic:

### Turnaround Time Modeling

Turnaround times are generated using a gamma distribution (shape=2, scale=8) to create a realistic right-skewed distribution. Additional business logic is applied:

* Complex services (Surgery, Chemo, Radiation): +30% TAT multiplier
* Urgent requests: -30% TAT multiplier for faster processing
* Denied requests: +20% TAT multiplier reflecting additional review time
* Base TAT range: 1-200 hours with realistic clustering around 24-72 hours

### Status Resolution Logic

The status resolution process follows realistic healthcare workflows:

* Initial status distribution: 70% approved, 10% denied, 15% pended, 5% received
* Pended requests: 60% resolve to approved/denied, 40% remain pended
* Appeal rates: 20% for denied requests, 5% for others
* Resubmission rate: 10% across all request types

## 3. Geographic Distribution Strategy

The dataset includes all 50 US states where Centene Corporation operates, with realistic health plan names and lines of business:

State selection is weighted towards larger population states (CA, TX, FL, NY, PA, IL, OH, GA, NC, MI) with 8% probability, while smaller states receive 5% probability. Each state is associated with authentic Centene health plan names and appropriate lines of business (Medicaid, Medicare, Marketplace).

## 4. Data Quality Assurance

Multiple quality assurance measures ensure data integrity and realism:

* Deterministic random seeds for reproducible results
* Comprehensive field validation and data type consistency
* Business rule enforcement (e.g., SLA calculations)
* Realistic timestamp generation with proper chronological ordering
* Appropriate null value handling for conditional fields

## 5. Data Visualization Strategy

Four key visualizations were created to demonstrate analytical capabilities:

### Status Distribution

Bar chart showing the distribution of prior authorization statuses, highlighting the high approval rate and low denial rate typical in healthcare systems.

### SLA Compliance by Urgency

Grouped bar chart comparing SLA compliance rates between urgent (24-hour SLA) and non-urgent (72-hour SLA) requests, demonstrating operational efficiency.

### Top Denial Reasons

Horizontal bar chart identifying the most common reasons for prior authorization denials, providing insights for process improvement.

### Turnaround Time Distribution

Histogram with log scale showing the distribution of processing times, with median line highlighting typical processing duration.

## 6. Technical Implementation

The solution is implemented as a single Python script using only standard libraries and common data science packages:

* Python 3.x with pandas for data manipulation
* NumPy for statistical modeling and random generation
* Matplotlib for data visualization
* python-docx for document generation
* Deterministic random number generation for reproducibility

## 7. Interactive Dashboard Development

A comprehensive HTML dashboard consolidates all visualizations into a single, professional interface suitable for executive presentations:

* Responsive design with modern CSS styling
* Summary statistics cards highlighting key metrics
* Integrated chart display with professional formatting
* Clear data source attribution and confidentiality notices
* Mobile-friendly layout for accessibility

## 8. Business Value Demonstration

This synthetic dataset demonstrates several key technical capabilities:

* Advanced synthetic data generation for healthcare applications
* Statistical modeling of complex business processes
* Data visualization and dashboard development
* Geographic and demographic analysis capabilities
* SLA monitoring and performance analytics
* Process improvement insights through denial reason analysis

## 9. Results Summary

Dataset successfully generated with 100,000 records:

* Overall approval rate: 77.3%
* Overall SLA compliance: 98.3%
* Urgent request SLA compliance: 91.0%
* Non-urgent request SLA compliance: 99.6%
* Geographic coverage: All 50 US states
* Health plans represented: 76 unique plans
* Lines of business: Medicaid, Marketplace, Medicare

## 10. Conclusion

This synthetic EDI-278 dataset successfully demonstrates advanced data engineering capabilities while maintaining complete confidentiality and data safety. The comprehensive approach to data generation, statistical modeling, and visualization provides a realistic foundation for healthcare analytics demonstrations without compromising patient privacy.

The solution showcases expertise in healthcare data standards, statistical modeling, data visualization, and dashboard development - all critical skills for modern healthcare data engineering roles.