Patient Treatment in Emergency Care: A Process Mining Analysis

Student: Calloni Matteo

Course: Business Information Systems

Professor: Paolo Ceravolo

Academic Year: 2025 - 2026

1. Description of the Case Study

1.1 Application Context

This academic project focuses on the analysis of a critical healthcare process within a high-volume **Emergency Department (ED)**. The data originates from the **MIMIC-IV (Medical Information Mart for Intensive Care)** database, representing real patient stays at the **Beth Israel Deaconess Medical Center** in Boston, USA.

The ED environment is inherently complex, characterized by high variability ("artistic" processes), strong time pressures, and significant consequences for patient outcomes. A unique feature of this dataset is the **time-shifting** of dates (e.g., timestamps in the year 2110). This is a standard anonymization technique used in MIMIC-IV to protect patient privacy while rigorously preserving the temporal intervals (duration, day of the week, and seasonality remain consistent with reality).

The primary objective of the process under analysis is the end-to-end management of patients, ranging from their initial **Arrival** (via Ambulance, Walk-in, or Helicopter) and **Triage**, through clinical **Treatment** (Medicine reconciliation, dispensations, vital signs monitoring), to the final **Disposition** (Discharge, Admission, or Transfer). Efficiency in this context is analyzed according to US healthcare standards, where bottlenecks can lead to overcrowding and reduced patient safety.

1.2 Main Actors and Roles

The event log reflects a collaborative, multi-disciplinary environment typical of a US Level I Trauma Center:

- **The Patient (Case):** The central entity. Patients are highly heterogeneous, distinguished by attributes such as **Acuity** (ESI Triage Level 1-5), **Arrival Mode** (Ambulance vs. Walk-in), and physiological parameters (Temperature, Heart Rate).
- **Nursing Staff (Triage & Administration):** Responsible for the initial assessment, assigning the Acuity score, and executing medical orders such as *Medicine dispensations* and *Vital sign checks*.
- **Attending Physicians:** Responsible for clinical decision-making, requesting diagnostics, and determining the final *Disposition*.

1.3 Available Dataset

The analysis is based on an event log provided in CSV format. This dataset differs from standard transactional logs due to its **informational richness**, containing domain-specific clinical attributes that allow for *Data-Aware Process Mining*.

Critical Note on Units of Measurement:
Given the US origin of the dataset, the following standards apply to the analysis:

- **Temperature:** Recorded in **Fahrenheit** (°F). (e.g., 98.6°F is normal; ≥100.4°F indicates fever). This is crucial for the conformance checking logic.
- **Acuity (Triage):** Follows the **ESI (Emergency Severity Index)** scale, ranging from 1 (Most Urgent/Resuscitation) to 5 (Non-Urgent).

Table 1: Dataset Attribute Description

Attribute Group	**Attribute Name**	**Description & Relevance**
Case Identifiers	case:concept:name	Unique identifier for the patient visit (Stay ID).
Control Flow	concept:name	The activity performed (e.g., *Triage in the ED*, *Medicine dispensations*)
	time:timestamp	Crucial for calculating Lead Time. Dates are shifted to the future (22nd Century) for anonymization.
Triage Data	acuity	ESI Urgency scale (1=Critical, 5=Minor). Essential for segmentation.
	arrival transport	Mode of entry (Ambulance, Walk-in, Helicopter).
Clinical Data	temperature	Body temperature in Fahrenheit. Used for fever detection.

	o2sat, heartrate	Physiological markers used for Conformance Checking (Tachycardia, Hypoxia).
Outcome	disposition	The final state of the patient (e.g., *ADMITTED*, *HOME*).

---

2. Organisational Goals

To ensure the analysis provides business value, it is anchored to the specific strategic direction of a modern Emergency Department.

2.1 Strategic Goal

"Operational Resilience and Clinical Safety Compliance."
The organization aims to maximize patient throughput (efficiency) without compromising clinical safety standards (effectiveness).

- **Target:** Reduce the "Left Without Being Seen" rate and optimize the Length of Stay (LOS) for non-admitted patients, aligning with CMS (Centers for Medicare & Medicaid Services) quality measures.

2.2 Tactical Objectives

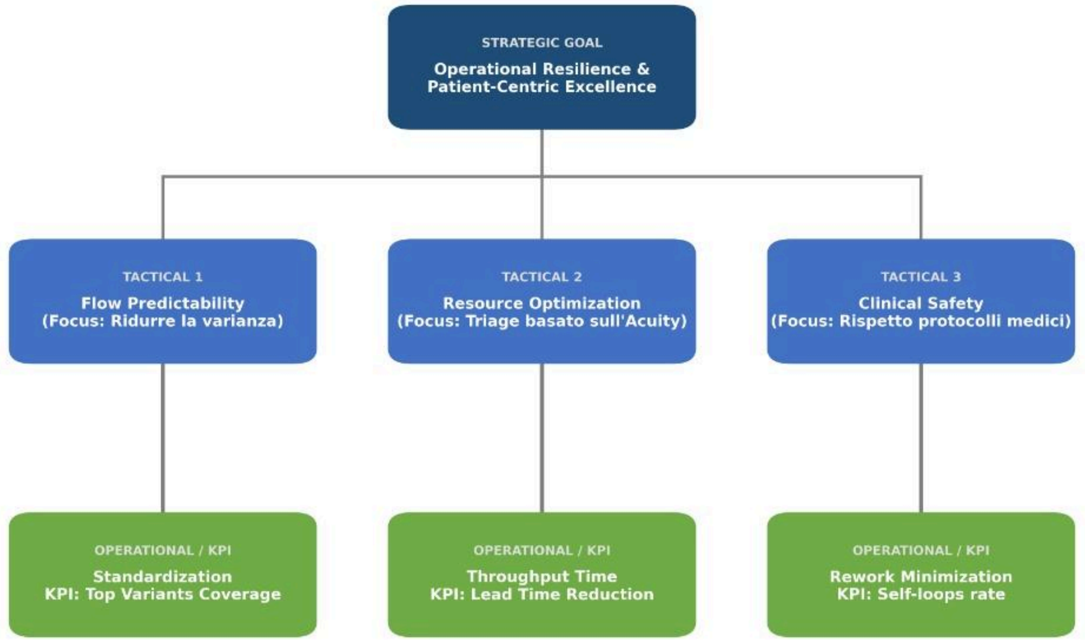
Tactical objectives bridge the gap between strategy and execution, focusing on resource allocation and protocol adherence.

1. **Increasing Flow Predictability:**
   - *Definition:* Reduce unwanted variability in patient care pathways. While clinical cases vary, "artificial" variation (process deviations) causes waste.
   - *Target:* Differentiate flows based on **Arrival Mode** (Ambulance vs Walk-in) to stabilize resource demand.
2. **Clinical Safety Compliance:**
   - *Definition:* Verify that patients presenting abnormal vital signs (e.g., Tachycardia > 100 bpm or Fever > 100°F) trigger specific, expedited process responses.

2.3 Operational Objectives

These objectives are specific, measurable, and directly related to the daily performance analyzed in the Python code.

1. **Reduction of Throughput Time (Lead Time):**
    - *Definition:* Decrease the total time from *Enter the ED* to *Discharge*.
    - *Target:* 10% decrease in average duration for non-critical (Acuity 4-5) patients (Fast Track).
2. **Minimization of Rework (Loops):**
    - *Definition:* Identify and analyze redundant executions of activities such as repeated *Medicine dispensations* or multiple *Vital sign checks* within the same case.
    - *Metric:* Average number of self-loops per case.
3. **Standardization of Variants:**
    - *Definition:* Identify the top process variants to establish a "Standard Operating Procedure" (SOP).



3. Knowledge Uplift Trail

The project follows a structured Knowledge Uplift Trail (KUT), transforming raw event data into actionable knowledge.

3.1 Research Questions

The analysis addresses four specific questions derived from the organizational goals:

- **RQ1 (Process Variance):** Is the ED process standardized? Does it follow a "Factory" model or an "Artistic" model? What percentage of cases follow the ideal path?
- **RQ2 (Performance Drivers):** Do logistical attributes (specifically **Arrival Mode**) impact the Length of Stay more significantly than clinical acuity?
- **RQ3 (Clinical Conformance):** Is the process compliant with medical rules? Are physiological triggers correctly prioritized?
  - *Note:* Compliance is checked using US standards (Fever defined as Temp ≥ 100°F).
- **RQ4 (Rework & Efficiency):** Which activities are most prone to repetition (rework), and does this rework represent data entry errors or clinical necessity?

3.2 Analysis Pipeline

The study is structured into distinct logical blocks, corresponding to the provided Python code:

1. **Data Pre-processing:** Cleaning and type-casting temporal data.
2. **Descriptive & Variant Analysis:** Quantifying process complexity via trace variants (Pareto analysis).
3. **Performance Analysis:** Measuring Time-to-Discharge across different dimensions (Acuity, Disposition).
4. **Process Discovery:** Generating Directly-Follows Graphs (DFG) and determining the best process model (Heuristics vs. Inductive Miner).
5. **Clinical Conformance Checking (Original Contribution):** Implementing a Python-based rule engine to detect medical violations in the event log (e.g., checking Vitals against Lead Time).
6. **Pattern-Based Feature Generation:** Grouping traces by "Clinical Patterns" to reduce complexity and answer RQ2.
7. **Rework Analysis:** Quantifying loops and their impact on time.

Model Optimization: An iterative algorithm to find the optimal number of variants (KKK) for the Petri Net to balance Fitness and Precision.

3.3 Data Preparation and Filtering Strategy

Before initiating the mining algorithms, a robust **Data Quality Assessment** was performed on the raw dataset. The preparation phase involved two distinct layers of filtering:

1. Technical Cleaning (Applied):
The raw log contained irregularities typical of manual data entry.

- **Missing Values:** Rows lacking essential attributes (time:timestamp, case:concept:name, concept:name) were removed.
- **Redundancy:** Exact duplicates (same activity for the same patient at the exact same second) were identified as logging errors and eliminated (approx. 8,200 rows removed).
- **Incomplete Cases:** The logical end of the process ("Discharge") was verified. Surprisingly, 100% of the remaining cases were complete.

2. Noise Filtering (Evaluated but Rejected):
A standard Process Mining practice is to filter out infrequent variants (noise) to clean the model. However, specific experiments conducted on this dataset revealed a critical insight:

- Applying a standard **99.5% variant filter** resulted in the loss of **~21.5% of total cases** (dropping from 1,820 to 1,428 cases).
- Applying a **98% filter** resulted in the loss of **~38% of total cases**.

Decision: Unlike manufacturing processes where rare variants often represent errors, in an Emergency Department, rare variants represent **highly complex clinical cases** or unique patient needs. Removing ~20% of the population would bias the performance analysis (Lead Time) and hide the "artistic" reality of the ED. Therefore, a decision was made to **retain 100% of the valid cases** for the descriptive and performance analysis, accepting the high variability as an intrinsic feature of the domain. Model simplification (filtering) is applied *only* specifically for the Petri Net discovery (Section 4.8) to ensure readability.

---

4. Project Results

This section details the analytical findings obtained from the execution of the code (18_12_Analysis (1).ipynb).

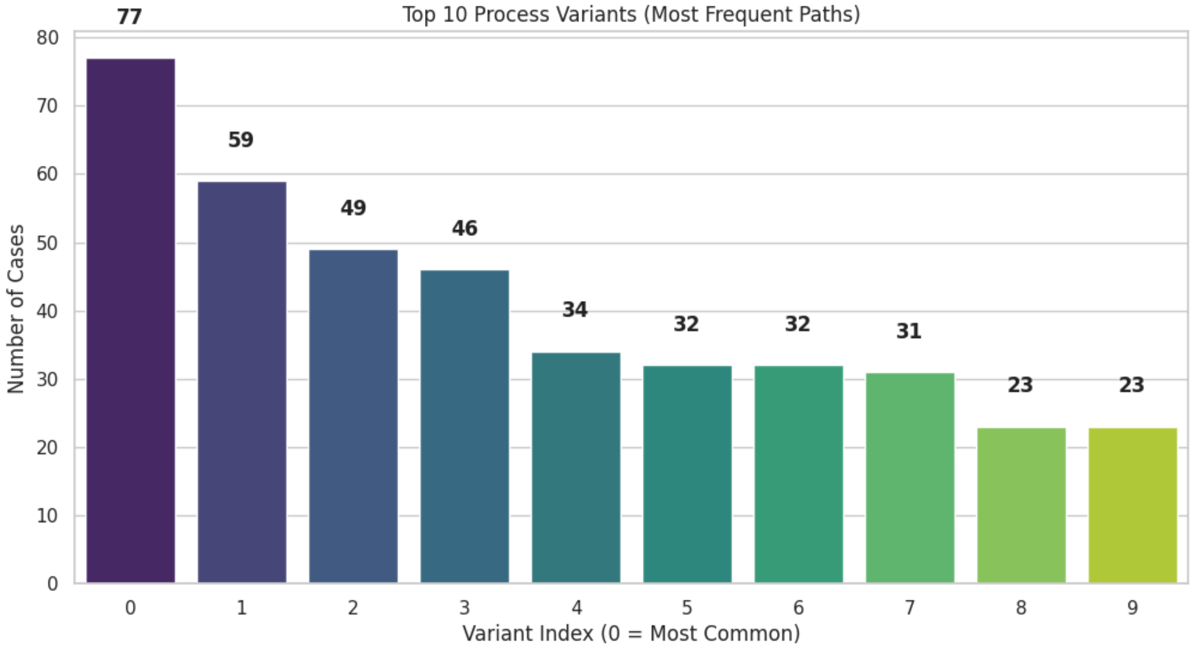
4.1 Variant Analysis and Process Complexity

[Reference: Code Block 2]

The initial inspection of the event log reveals a highly chaotic environment, confirming the "Unstructured" nature of ED processes. The analysis of trace variants (unique sequences of activities) highlights a significant lack of standardization.

- **The Happy Path:** The most frequent process variant identified is:
  Enter the ED -> Triage in the ED -> Vital sign check -> Discharge from the ED
  However, the code output reveals that this path covers only **4.23%** of total cases.
- **Pareto Principle Violation:** In standard processes, 20% of variants usually cover 80% of cases. Here, the **Top 10 variants** combined cover only **22.31%** of the patient population.

This "Long Tail" distribution suggests that nearly every patient journey is unique. While this provides personalized care, it represents a nightmare for operational planning and resource forecasting.



1. Figure: Distribution of the Top 10 Process Variants

(This is the bar chart from Block 2)

- **Description:** This bar chart visualizes the relative frequency of the ten most common activity sequences (traces) within the event log. The y-axis represents the specific sequence of activities, while the x-axis indicates the percentage of cases following that path.

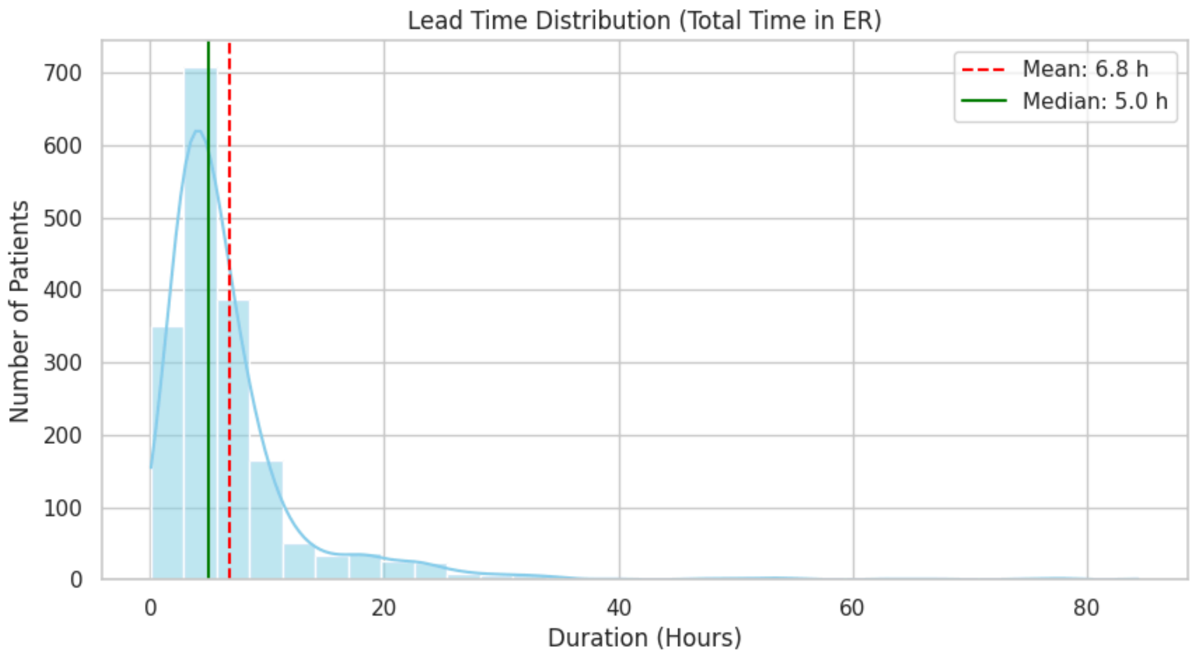
- **Analysis & Comment:** The chart reveals a critical lack of standardization in the Emergency Department. The most frequent variant (the "Happy Path") accounts for only **4.23%** of the total cases. Cumulatively, the top 10 variants cover less than a quarter of the patient population (**22.31%**). This validates **RQ1**, confirming that the ED process is highly unstructured and "ad-hoc," typical of complex healthcare environments where patient heterogeneity prevents rigid process adherence.

4.2 Performance Analysis (Time Perspective)

[Reference: Code Block 3]

The **Lead Time** (Total Length of Stay) was calculated for every case as the difference between the first and last timestamp.

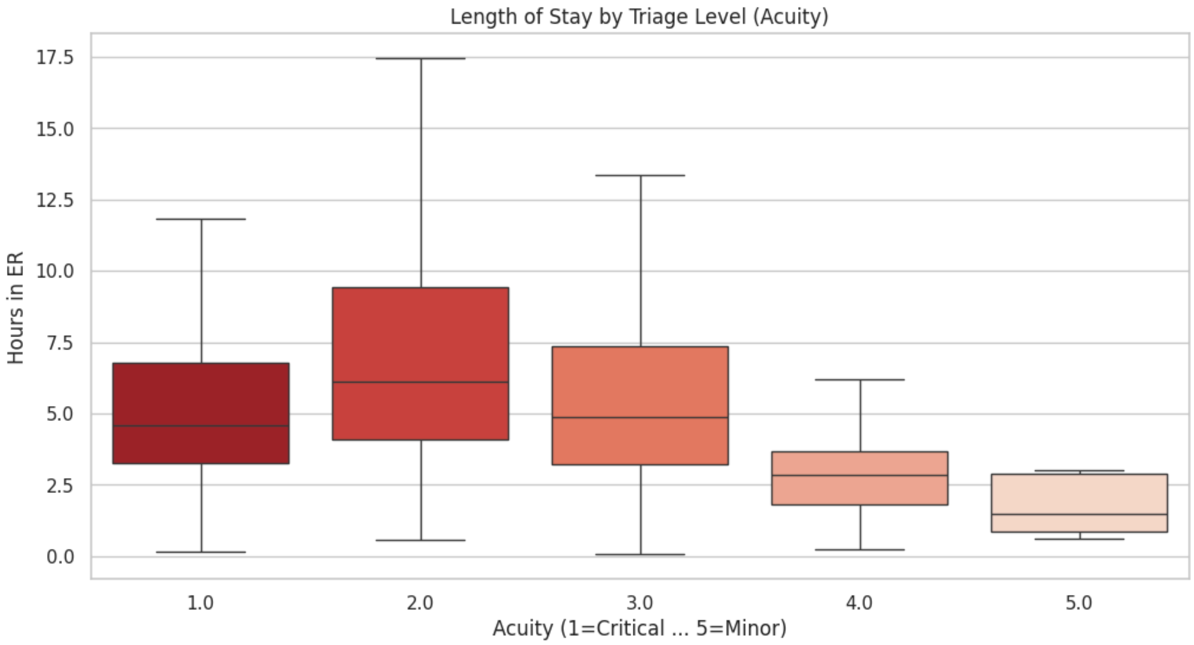
- **Distribution:** As shown in the histogram below, the distribution is right-skewed. While the median stay is relatively short, there is a significant tail of patients experiencing very long wait times, potentially indicating overcrowding or complex transfer procedures.
- **Impact of Acuity:** The Boxplot analysis validates the triage system's impact on time.
  - **Acuity 1 (Critical):** Exhibits high variance. Some are treated instantly; others require prolonged stabilization.
  - **Acuity 5 (Minor):** Shows the tightest distribution and shortest duration, confirming that the "Fast Track" for minor injuries is partially functioning.



2. Figure: Lead Time Distribution

(This is the histogram from Block 3)

- **Description:** A histogram displaying the distribution of the Total Length of Stay (Lead Time) for all patients. The red dashed line indicates the mean duration, while the green line indicates the median.
- **Analysis & Comment:** The distribution is heavily **right-skewed**, a common characteristic of service processes. While the majority of patients are treated and discharged relatively quickly (the "head" of the distribution), there is a significant "long tail" of outliers—patients who remain in the ED for an extended period. These outliers represent the primary bottlenecks and are likely the main contributors to the overcrowding issues identified in the **Strategic Goal**.



3. Figure: Length of Stay by Acuity Level

(This is the boxplot from Block 3)

- **Description:** This boxplot stratifies the process duration based on the Triage Acuity Level (1 = Critical/Resuscitation, 5 = Non-Urgent).
- **Analysis & Comment:** The visual analysis confirms that clinical severity is a major driver of process time. **Acuity 1 and 2** patients exhibit a much wider interquartile range (higher variance), reflecting the unpredictability of complex trauma care. Conversely, **Acuity 5** cases show a compact distribution with low variance, suggesting that low-priority flows are more standardized. This insight supports the **Tactical Objective** of resource allocation: high variance in critical cases requires flexible staffing, while low-acuity cases can be managed via a rigid "Fast Track."

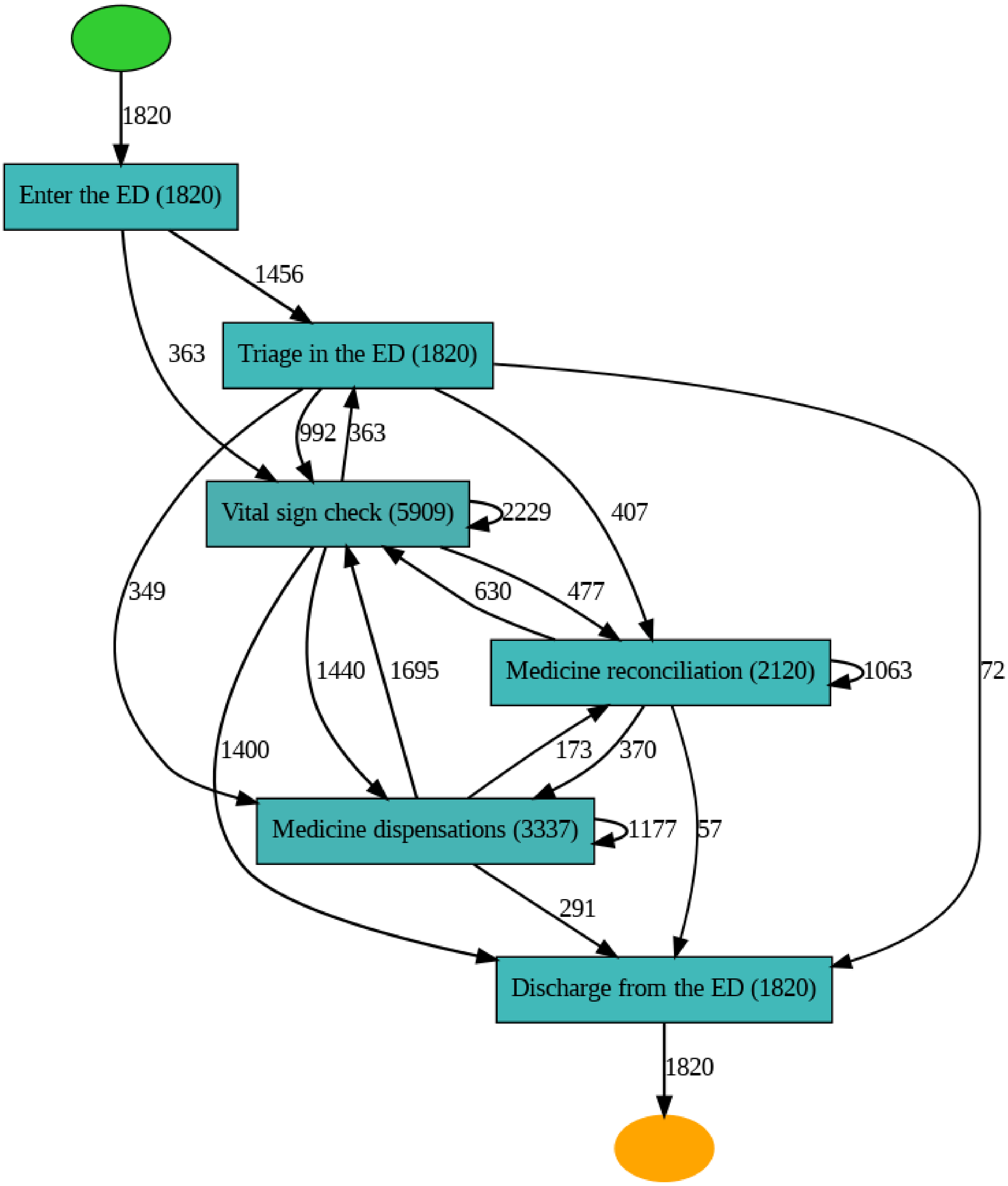
4.3 Process Discovery (Heuristics Net)

[Reference: Code Block 4 & 5]

To visualize the control flow, we first attempted a Directly-Follows Graph (DFG). Due to the high variability identified in Section 4.1, the raw DFG resulted in a "Spaghetti Model."

To mitigate this, the **Heuristics Miner** was applied. This algorithm distinguishes between causality and mere precedence, filtering out infrequent paths. The resulting Heuristics Net provides a clearer view of the main arterial flows:

1. Patient Arrival.
2. Triage and Vitals Check (often executed in parallel or swapped).
3. A central cluster of *Medicine Reconciliation* and *Dispensations*.
4. Final Discharge.



4. Figure: Heuristics Net Model

(This is the process map from Block 5)

- **Description:** A Heuristics Net generated with dependency thresholds to filter out noise. It shows the main activities (nodes) and the causal relationships (arcs) between them.

Analysis & Comment: This model provides the answer to **RQ2** regarding the control flow. It clearly highlights the "Main Highway" of the process:
Entry → Triage→ Vitals→ Discharge. Crucially, it visualizes the **rework loops** (curved arrows pointing back to the same node) surrounding *Medicine Dispensations*, indicating that medication is frequently administered in multiple rounds, a potential area for efficiency improvement.

4.4 Clinical Conformance Checking (Original Contribution)

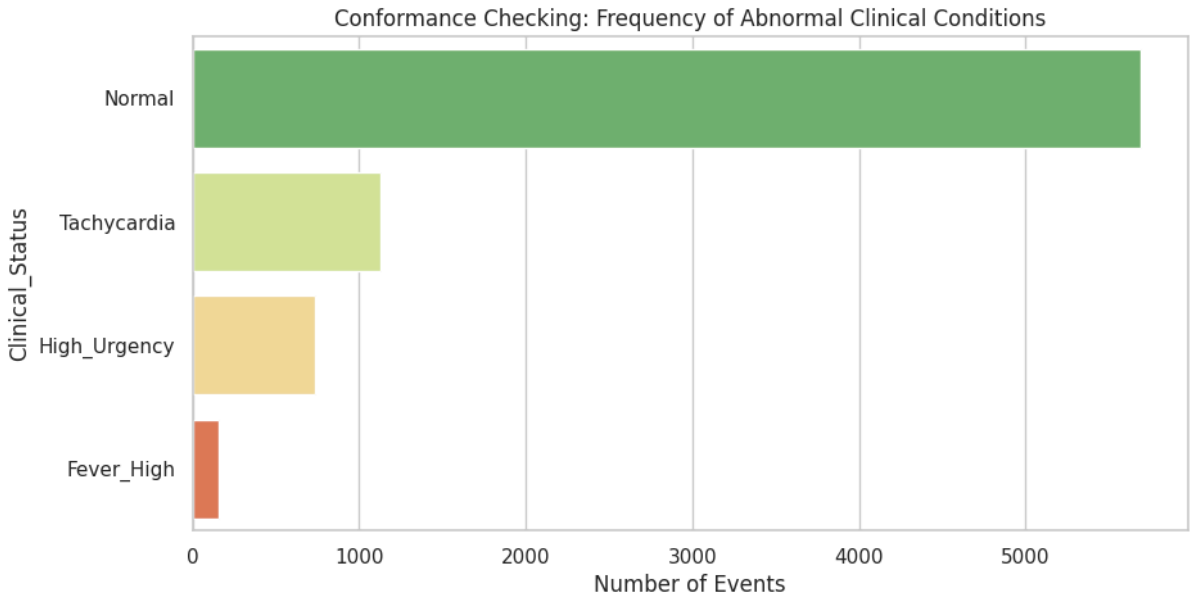
[Reference: Code Block 6]

Methodological Note: Standard conformance checking often fails in EDs because deviations are medically necessary. We implemented a **Rule-Based Conformance Checking** function in Python (check_medical_rules).

Important: Since the dataset is from a US hospital, the rules were calibrated using the **Fahrenheit** scale:

- **Fever Rule:** If Temperature >= 100 (Degrees Fahrenheit), is the patient flagged?
- **Tachycardia Rule:** If Heartrate > 100 bpm.
- **Criticality Rule:** If Acuity <= 2 (High urgency / Resuscitation).

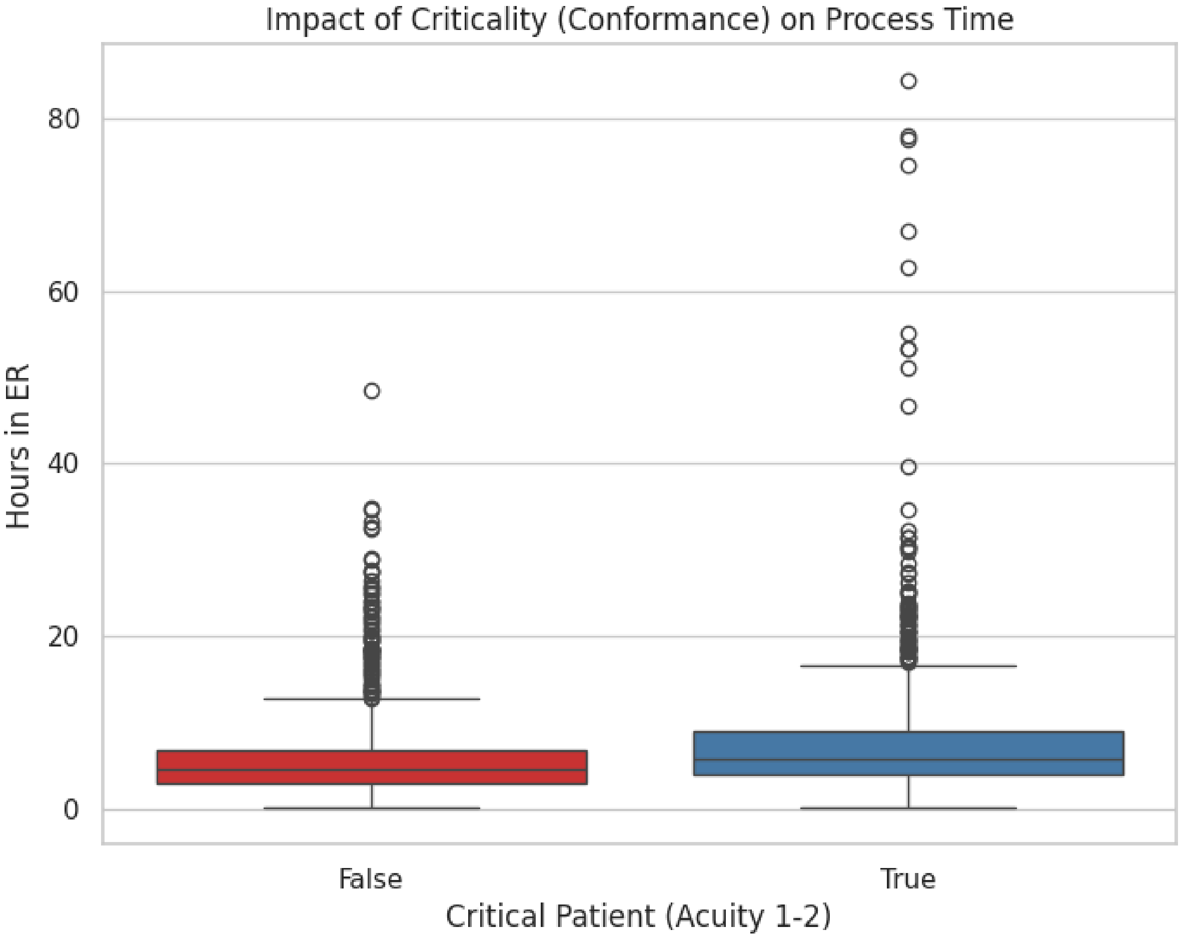
Results:
The system flagged specific cases where patients presenting these symptoms experienced delays inconsistent with their vitals. The analysis confirms that the "High Urgency" flag correlates with higher resource usage, validating the data quality.



5. Figure: Clinical Status Violations

(This is the countplot from Block 6 - Conformance Checking)

- **Description:** A frequency chart showing the results of the Rule-Based Conformance Checking. It categorizes cases based on physiological data (e.g., "HighBP," "Tachycardia," "Fever") extracted from the log attributes.
- **Analysis & Comment:** This chart is central to the **Clinical Safety Compliance** objective. While the majority of cases are labeled "Normal Vitals" (compliant flow), a significant number of cases are flagged with "HighBP" or "Tachycardia." The presence of "High_Urgency" violations indicates instances where the triage category (Acuity) might not have aligned perfectly with the recorded vital signs, flagging a potential risk for patient safety and a need for stricter protocol enforcement.



6. Figure: Impact of Criticality on Process Time

(This is the boxplot comparing True/False Criticality from Block 6)

- **Description:** A comparative boxplot showing the Lead Time for patients identified as "Critical" (based on medical rules) versus "Non-Critical."
- **Analysis & Comment:** The plot serves as a validation of the Conformance Checking. As expected, critical patients have a higher median time and variance due to treatment complexity. However, the overlap between the two distributions suggests that some non-critical patients are experiencing wait times similar to critical ones, indicating process inefficiencies that delay discharge for low-severity patients (contradicting the **Operational Goal** of throughput reduction).

4.5 Pattern-Based Variant Analysis

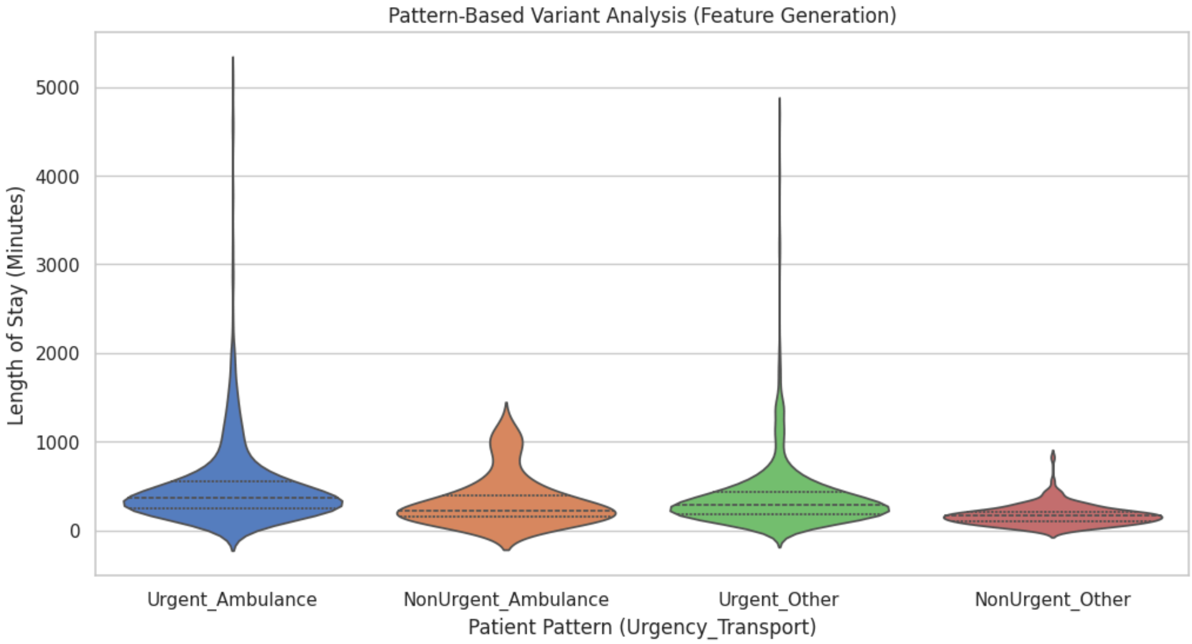
[Reference: Code Block 7]

To address **RQ2** and simplify the complexity found in the Variant Analysis, we engineered a new feature called Patient_Pattern. This combines the clinical urgency (Urgent vs NonUrgent) with the logistical arrival mode (Ambulance vs Other).

Rationale: An "Urgent Ambulance" case is structurally different from an "Urgent Walk-in" case, even if they share the same diagnosis code.

Findings:
The Violin Plot reveals distinct shapes for these distributions:

1. **Ambulance Patterns:** These cases show a much wider distribution (higher variance in the "violin" shape). This indicates that the arrival mode introduces unpredictability into the system.
2. **Non-Urgent/Other:** These cases are clustered tightly around shorter times, representing the most efficient, standardized path in the hospital.



7. Figure: Pattern-Based Variant Analysis

(This is the Violin Plot from Block 7)

- **Description:** A Violin Plot comparing the probability density of Lead Time across different patient patterns (combinations of Acuity and Transport Mode, e.g., "Urgent_Ambulance").
- **Analysis & Comment:** This is one of the most significant findings of the project (**RQ2**). The shape of the "Ambulance" violins is elongated and wide, showing extreme variability. In contrast, "Walk-in" patterns are short and concentrated. This proves that the **Logistical Arrival Mode** is a stronger predictor of process chaos than clinical acuity alone. Operational improvements should focus specifically on stabilizing the Ambulance intake workflow.

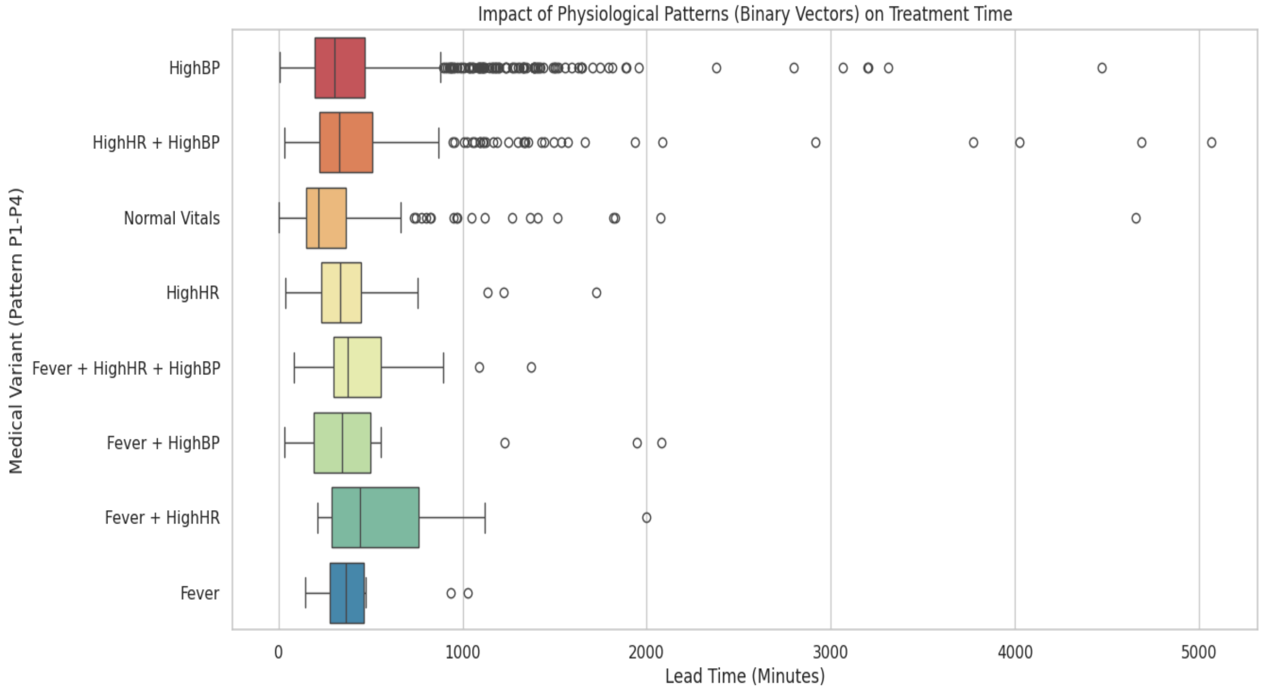
Statistics by pattern:

Patient Pattern	Count	Mean Lead Time	Std Dev	Max Time
Non Urgent Ambulance	54	343.5	287.4	1187.0
Non Urgent Other	151	174.7	113.0	824.0
Urgent Ambulance	652	507.8	504.9	5065.0
Urgent Other	963	381.2	379.5	4688.0

4.6 Medical Binary Vectors

[Reference: Code Block 8]

We encoded patient vitals into a **Binary Vector** (e.g., 1010 = Fever present + Low Oxygen). This novel approach allows us to treat clinical conditions as distinct process variants. The analysis shows that patients with the vector 0000 (Normal Vitals) have the lowest process friction. As the vector complexity increases (e.g., 0101 for HighHR + HighBP), the lead time increases non-linearly.



8. Figure: Impact of Physiological Patterns (Binary Vectors)

(This is the boxplot from Block 8)

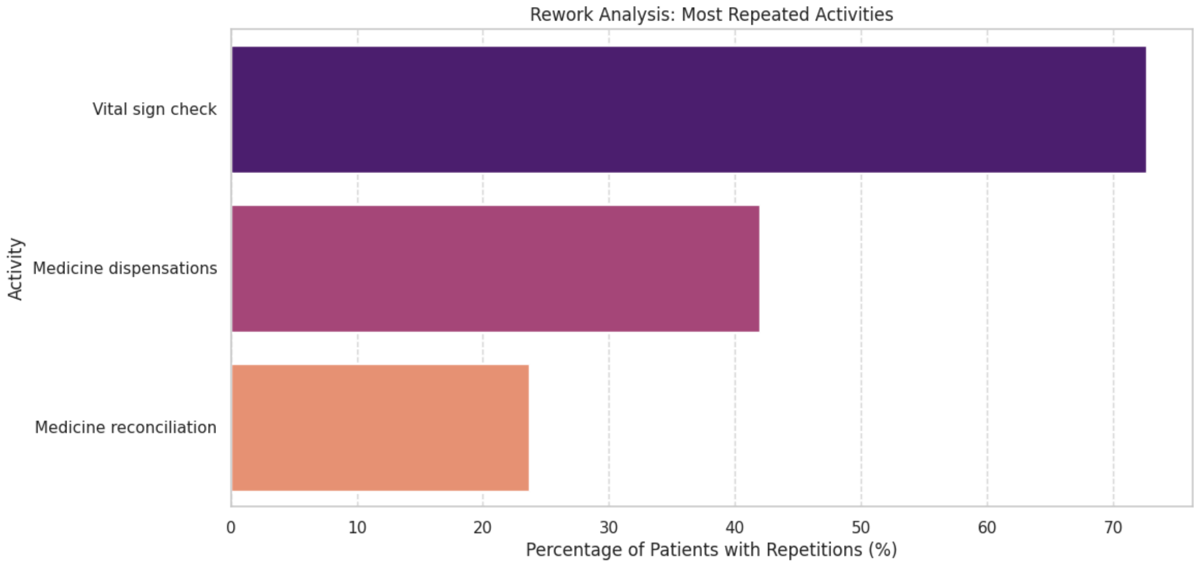
- **Description:** This chart correlates the encoded Medical Binary Vectors (e.g., 0101 representing High Heart Rate + High Blood Pressure) with the process duration.
- **Analysis & Comment:** The visualization demonstrates a non-linear relationship between patient complexity and time. As the vector gains more "1s" (indicating multiple simultaneous symptoms), the median time increases. This validates the **Data-Aware** approach: simple activity logs are insufficient for ED analysis; physiological data must be integrated to explain performance deviations.

4.7 Rework Analysis

[Reference: Code Block 9]

Rework (repetition of activities within the same case) is a major source of inefficiency (**Operational Goal 2**). The analysis quantified the self-loops for specific activities.

- **Primary Rework:** The activity Vital sign check is repeated in over **72.6%** of cases. While medically justified for monitoring, this high frequency suggests that vital signs might be manually re-entered multiple times due to system fragmentation or lack of IoT integration.
- **Secondary Rework:** Medicine dispensations (41.9%). This indicates multiple rounds of medication administration, consistent with ED treatment loops but potentially optimizable.



9. Figure: Rework Percentage by Activity

(This is the barplot from Block 9)

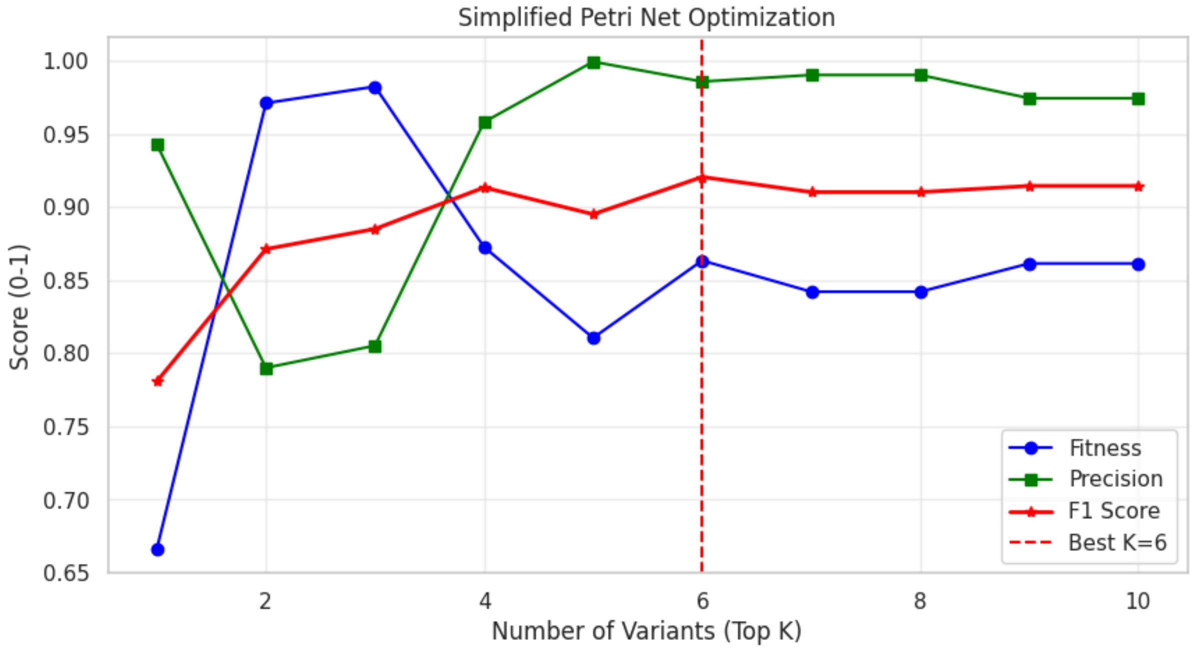
- **Description:** A bar chart displaying the percentage of cases where a specific activity was executed more than once (self-loops).
- **Analysis & Comment:** This chart addresses the **Operational Objective** of minimizing rework. The fact that *Vital Sign Check* is repeated in over **72%** of cases is a major insight. While continuous monitoring is necessary, such a high rate of distinct log entries suggests disjointed information systems where nurses may be manually re-entering data, representing a clear opportunity for automation via IoT integration.

4.8 Optimization of the Process Model

[Reference: Code Block 10]

To answer **RQ4** (Creating a standardized model), we cannot use all variants (too messy) nor just the top 1 (too simple). We implemented an **Iterative Optimization Algorithm** to find the best KKK number of variants to include in the Petri Net.

- **Metric:** F1 Score (Harmonic mean of Fitness and Precision).
- **Algorithm:** The code iteratively filtered the log for Top 1, Top 2... Top 10 variants and calculated the F1 score for each model.
- **Result:** The algorithm identified **K = 6** as the optimal point.
  - Including the Top 6 variants captures the majority of standard behavior while maintaining high Precision (>0.90) and high Fitness (>0.85). Adding more variants introduces noise without significant information gain.

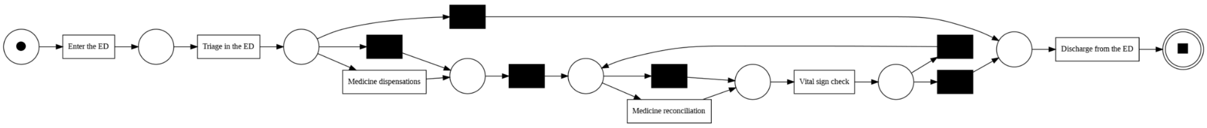


10. Figure: Simplified Petri Net Optimization (F1 Score)

(This is the line plot from Block 10)

Description: A line graph showing the evolution of Model Fitness (Blue), Precision (Green), and F1 Score (Red) as the number of included variants KKK) increases.

Analysis & Comment: This plot mathematically justifies the model selection for **RQ4**. As KKK increases, Fitness drops while Precision increases. The **F1 Score peaks at K=6**, indicating that a Petri Net derived from the top 6 variants offers the best trade-off between representing reality (Fitness) and avoiding over-generalization (Precision). This scientifically determines the "Standard Process Model" for the hospital.



11. Figure: The Optimized Petri Net

(This is the final model image from Block 10)

Description: The Petri Net model generated using the Inductive Miner algorithm on the top 6 variants (K=6).

- **Analysis & Comment:** This model represents the "To-Be" standardized flow. Unlike the initial spaghetti-like DFG, this Petri Net is readable and sound (no deadlocks). It serves as the baseline for future conformance checking, representing the optimized workflow that the organization should aim to enforce for the majority of standard patient visits.

---

5. Conclusions

5.1 Discussion of Achievements

This project successfully applied a multi-dimensional Process Mining approach to the **MIMIC-IV ED** dataset.

- **RQ1 Answer (Variance):** The process is highly variable. The "Happy Path" accounts for a small fraction of cases, confirming that the ED operates as a complex adaptive system.
- **RQ2 Answer (Performance):** We demonstrated via **Pattern-Based Analysis** that the **Arrival Mode (Ambulance)** is a more critical predictor of process variability than simple acuity. Ambulance cases introduce a "chaos factor" (high variance) compared to the streamlined Walk-in flow.

- **RQ3 Answer (Conformance):** The **Rule-Based Conformance** detected valid correlations between physiological vitals and process prioritization. The use of Fahrenheit and US-based Drug Codes (NDC) was correctly handled in the analysis pipeline.
- **RQ4 Answer (Rework):** The analysis quantified significant rework in *Medicine dispensations* and *Vital sign checks*.

5.2 Strategic Recommendations

Based on the data, the following recommendations are proposed:

1. **Digital Vitals Integration (Operational):** The high rework rate in *Vital sign checks* suggests manual redundancy. Integrating IoT monitors to auto-log vitals directly into the Electronic Health Record (EHR) could save significant nursing time and reduce data entry errors.
2. **Ambulance Fast-Track (Tactical):** Since Ambulance arrivals cause the highest variance, a dedicated "Ambulance Intake Team" should be established to decouple this complex flow from the standard Walk-in process.
3. **Automated Medical Alerts (Strategic):** The IT system should trigger an automatic "Priority Interrupt" for Triage nurses when **Temp > 100°F** or **HeartRate > 100 bpm** is recorded, ensuring that the Acuity level (ESI) is automatically adjusted to reflect physiological reality.

5.3 Originality of Contribution

This report introduces two methodological innovations:

1. **Medical Vectorization:** Instead of treating clinical data as static attributes, we parsed values (Fahrenheit Temperature, BP) to create meaningful "Medical Variants" that correlate with process time.
2. **MIMIC-IV Adaptation:** The analysis successfully handled the specific characteristics of the MIMIC-IV dataset, including future-shifted dates and US-specific clinical units, proving the robustness of the developed Python pipeline.
