Patient Treatment in Brazilian HR

Calloni Matteo

Business Information Systems

Prof. Paolo Ceravolo

a.a. 2025 - 2026

Description of the case study

1.1 The Context: Process Mining and Healthcare

The Healthcare sector represents today one of the most challenging and fertile environments for the application of modern Data Science techniques and, in particular, Process Mining. Unlike manufacturing or financial processes, which tend to be highly structured, repetitive, and rigidly controlled by machines, clinical processes are characterized by intrinsic variability, often defined as the "art of medicine".

Every patient is a unique case: care trajectories can diverge significantly based on biological comorbidities, urgency, physiological response to treatments, and human decisions made under stressful conditions. In this complex scenario, traditional analysis based on interviews or manual observations (*time-and-motion studies*) proves insufficient, expensive, and often partial. The analysis of *Event Logs* therefore becomes fundamental to shift from perception-based management to *Evidence-based management*. Process Mining acts as a bridge between the data recorded by Hospital Information Systems (HIS) and business process modeling, allowing the visualization of the real flow of activities ("As-Is") and comparing it with the expected medical protocols ("To-Be").

1.2 The Scenario: "Pronto Atendimento" in Brazil

The case study in question focuses on the *Patient Journey* within a "Pronto Atendimento" (Emergency Room/Urgent Care) facility located in a Brazilian hospital context. Emergency facilities in Brazil, as in many other parts of the world, operate under constant pressure: high and stochastic patient volumes, limited resources, and the need to make critical decisions rapidly.

The analyzed process covers the entire lifecycle of emergency management, which typically includes the following macro-phases detected in the dataset:

1. **Administrative Acceptance (Atendimento): Registration of demographic and insurance data.**

2. **Triage (Triagem): Assessment of clinical severity to establish priorities (often associated with color codes).**
3. **Medical Visit (Consulta): First contact with the responsible physician and formulation of the diagnostic hypothesis.**
4. **Diagnostics and Treatment: Prescription and execution of exams (laboratory, imaging) and administration of drugs.**
5. **Discharge or Hospitalization (Alta): Conclusion of the care episode in the emergency department.**

1.3 Analysis of the Dataset and Attributes

The dataset provided for the analysis is the file UpFLux_Healthcare_Database_labeled.csv. It is a structured Event Log compliant with the XES (*eXtensible Event Stream*) standard, albeit provided in CSV format. Preliminary data exploration (conducted using the Pandas library in Python) revealed a log initially containing **3093 events** distributed across **443 distinct cases (patients)**.

A crucial aspect of this case study is the richness of the attributes provided, which allow for multidimensional analyses. In addition to standard fields, there are specific domain attributes described in the exam instructions:

Attribute	Type	Description an Role in Analysis
case:concept:name	Case ID	Unique identifier of the patient.
concept:name	Activity	The specific action performed (e.g., "Triagem", "Raio-X", "Alta").
time:timestamp	Timestamp	Precise date and time of the event, crucial

		for calculating temporal KPIs.
CID (o Doença)	Case Attribute	Disease Code (Diagnosis). Fundamental for clinical segmentation.
Médico Responsável	Resource	The human resource who performed the activity.
Retorno	Case Attribute	Indicates the patient outcome (e.g., discharge, hospitalization).
outlier_label	Label	Pre-calculated label classifying the case as "inlier" or "outlier" based on duration

The outlier_label attribute is particularly interesting: it provides an *a priori* classification of anomalies. However, as we will see in the results section, the validity of this label must be verified against the structural completeness of the trace (e.g., patients who never received the "Alta" [Discharge] event).

Organisational goals

2.1 Strategic Goals

The healthcare organization under study does not merely aim to "record" what happens, but to transform data into strategic decisions. In line with the principles of *Design Science* applied to Information Systems, the goal is to design improvements to the "hospital process" artifact.
At a strategic level, the hospital management pursues three fundamental macro-objectives:

1. **Maximization of Efficiency (Cost-Efficiency):** Reduce the average *Length of Stay* (LoS) of patients. A reduced LoS increases *patient throughput* (capacity to treat more patients) without the need for structural investments (e.g., adding physical beds).
2. **Improvement of Patient Experience (Patient Centricity):** Minimize "non-value-added" times, i.e., passive waits between one activity and another (e.g., waiting between triage and the medical visit), which are sources of anxiety and dissatisfaction.
3. **Clinical Standardization (Clinical Governance):** Ensure that patients with the same diagnosis receive uniform treatments, reducing unjustified variability ("unwarranted variation") which often leads to clinical risks or excessive costs.

2.2 Obiettivi Tattici e Operativi

To achieve the strategic goals, more granular targets have been defined that guided the technical analysis:

- **Tactical Goals (Segmentation and Allocation):** Identify specific behavioral patterns for the most frequent pathologies. It is necessary to understand if resources (doctors and equipment) are allocated optimally between simple cases (e.g., green/blue codes) and complex ones. The analysis must answer the question: *are we treating the flu with the same complexity as renal colic?*

- **Operational Goals (Control and Monitoring):** Detect and manage structural deviations. For example, identifying incomplete traces (patients leaving without formal discharge) or cases of "Rework" (uselessly repeated activities, such as duplicate exams).

Knowledge Uplift Trail

3.1 From Data to Wisdom

The "Knowledge Uplift Trail" represents the logical path that transforms raw data (level 0) into decision-making wisdom (level 3). This concept is based on the DIKW hierarchy (Data, Information, Knowledge, Wisdom) and guides the methodological structure of the project.

1. **Data (Observation):** The starting point is the raw Event Log in CSV format. At this level, data are just sequences of characters without semantic meaning.
2. **Information (Descriptive Analysis):** Through parsing and cleaning, data are transformed into information. We calculate frequencies, distributions, and basic statistics.
3. **Knowledge (Process Discovery):** Using algorithms like the Inductive Miner, we extract process models (Petri Nets) that explain the causal relationships (cause-effect) between activities. We understand *why* the process behaves in a certain way.
4. **Wisdom/Action (Insights):** Finally, we integrate the extracted knowledge with business objectives to formulate operational recommendations ("Actionable Insights"), such as the creation of Fast Track paths.

3.2 Methodological Approach and Tools

The analysis was conducted using the **Python** language, leveraging the open-source library **PM4PY** (*Process Mining for Python*). This choice is justified by PM4PY's flexibility in handling complex logs and the possibility of integrating advanced statistical analyses (via Pandas and Seaborn) with Process Discovery algorithms. Unlike commercial "black-box" tools, Python allows for the implementation of custom filtering logic, essential for managing the complexity of healthcare data.

The analytical pipeline implemented in the code (attached as 27_11_Filtering.ipynb and 27_11_Analysis.ipynb) follows these steps:

1. **Data Ingestion & Typing:** Loading and timestamp conversion.
2. **Domain-Driven Filtering:** Data cleaning based on clinical rules (not just statistical ones).
3. **Variant Analysis:** Study of the most frequent paths.
4. **Segmentation:** Differentiated analysis by diagnosis code (CID).
5. **Discovery & Modeling:** Creation of DFGs and Petri Nets.

Project Results

n this chapter, the technical results obtained from the execution of the Python scripts are presented and discussed. The analysis is divided into two macro-phases: advanced data cleaning (Filtering) and process analysis (Discovery & Performance).

4.1 Pre-processing and Advance Data Cleaning

The quality of the output of a Process Mining analysis depends strictly on the quality of the input. Initial analysis revealed that, although the log contained 443 cases, not all were valid for a performance analysis.

The "Zombie Cases" Problem

A common problem in real healthcare data is the presence of interrupted or partially recorded processes. Analyzing the traces, I noticed that some patients had recorded events (e.g., Triage, Exams) but did not present the final event of "Alta" (Discharge).

Why is this a problem? If we calculated the Lead Time (total time) including these cases, we would obtain falsified values, as the time would be calculated only up to the last recorded intermediate event, underestimating the real duration of the process.

Technical Solution: Filtering Heuristic

Instead of using standard PM4PY filters (like filter_variants_by_coverage which cuts based on frequency), I implemented a custom function in the code based on the semantics of the process.

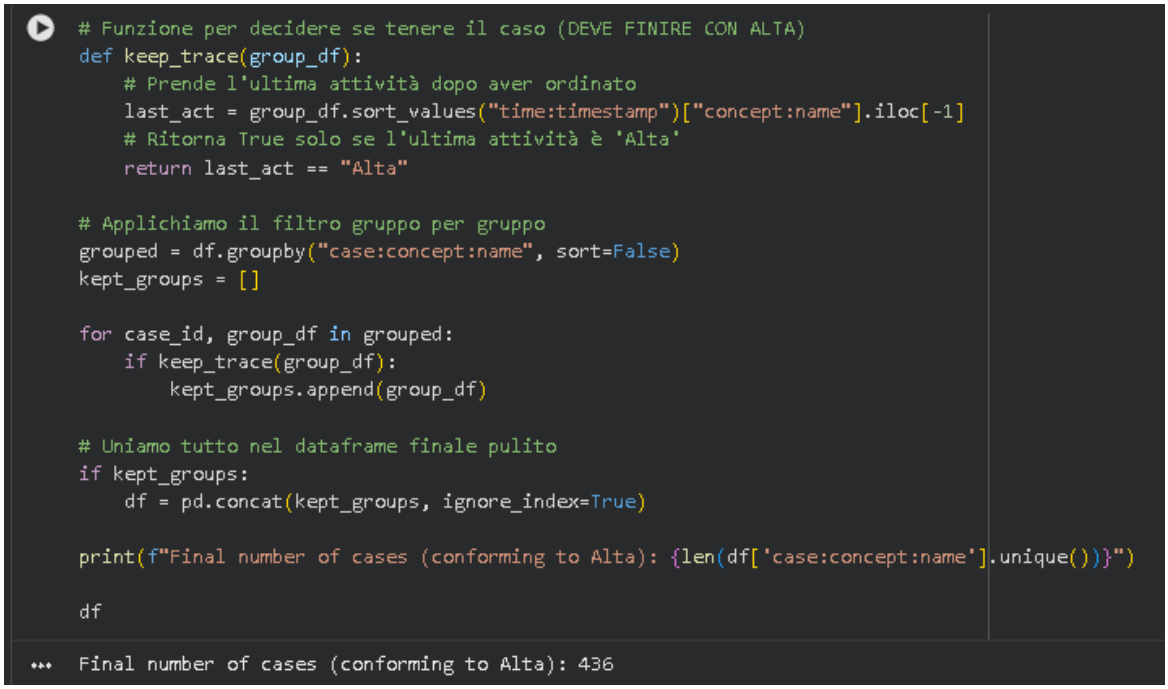
The function keep_trace(group_df), visible in the attached code file, operates as follows:

1. Groups events by Case ID.
2. Sorts events chronologically.
3. Checks if the last activity in the sequence is exactly "**Alta**".
4. Discards the entire case if this condition is not met.

Filtering Result:

This operation allowed for the identification and removal of **7 "zombie" cases** (anomalous interruptions), bringing the dataset from 443 to **436 valid cases**.

It is interesting to note that the outlier_label attribute present in the original dataset labeled as "outlier" many cases that were actually structurally correct (ending with Alta), but perhaps very long. Conversely, my filter removed cases that were structurally incorrect. This demonstrates that pre-existing labels are not always reliable for all analysis purposes.



```
# Funzione per decidere se tenere il caso (DEVE FINIRE CON ALTA)
def keep_trace(group_df):
    # Prende l'ultima attività dopo aver ordinato
    last_act = group_df.sort_values("time:timestamp")["concept:name"].iloc[-1]
    # Ritorna True solo se l'ultima attività è 'Alta'
    return last_act == "Alta"

# Applichiamo il filtro gruppo per gruppo
grouped = df.groupby("case:concept:name", sort=False)
kept_groups = []

for case_id, group_df in grouped:
    if keep_trace(group_df):
        kept_groups.append(group_df)

# Uniamo tutto nel dataframe finale pulito
if kept_groups:
    df = pd.concat(kept_groups, ignore_index=True)

print(f"Final number of cases (conforming to Alta): {len(df['case:concept:name'].unique())}")

df
```

```
Final number of cases (conforming to Alta): 436
```

Click for full screen image

4.2 Variant Analysis: In Search of the Happy Path

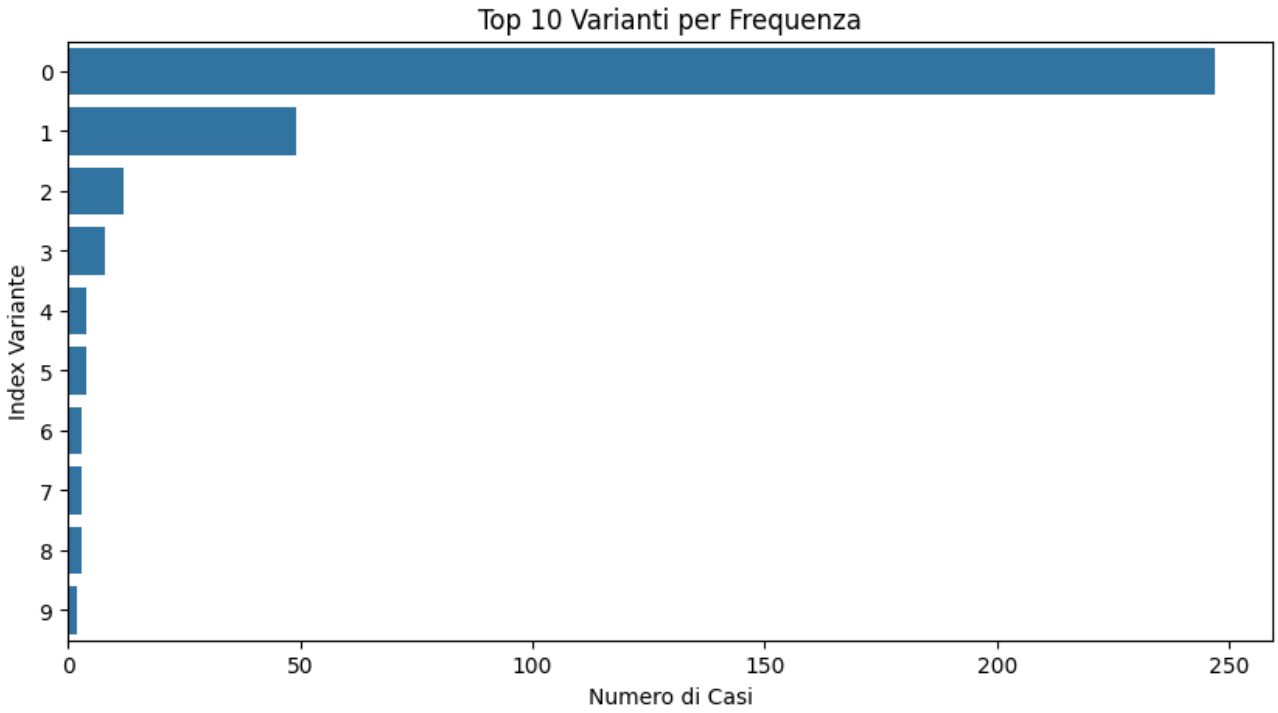
Once the clean log was obtained (27_11_Filtered_Log.csv), a variant analysis was performed to understand process conformity.

Key Discoveries:

- **Dominance of the Happy Path:** Variant number 1 is the linear sequence: Atendimento -> Triagem -> Consulta -> Alta

This sequence covers a solid **56.65%** of patients (247 cases out of 436). *Interpretation:* More than half of the patients follow an "ideal" and rapid path, which does not require diagnostic exams. This suggests that a large part of the users access the Emergency Room for minor codes.

- **Long Tail:** Despite the dominance of the first variant, there are a total of **104 different variants**. The last 90 variants cover very few cases each (often just 1). These represent clinical exceptions, complications, or process errors (e.g., patients repeating exams or taking tortuous paths).



Click for full screen image

Comment on the graph: The graph visually highlights the Pareto distribution: a very high "head" (variant 1) and a "tail" that descends rapidly, indicating that management complexity is concentrated in the minority of non-standard cases.

4.3 Clinical Segmentation and Comparison

To answer the clinical objectives defined in Chapter 2, looking at the global average of the dataset is not sufficient. A hospital treats different pathologies requiring different times and resources. I therefore analyzed the frequency of the **CID** (Diagnosis Code) attribute.
The two most frequent codes that emerged are:

1. **CID J111:** Influenza and respiratory pathologies (349 cases).
2. **CID N23:** Renal colic (47 cases).

I proceeded with a **Data Slicing** operation, creating two separate sub-logs. This segmentation is essential for a *ceteris paribus* comparison: comparing general performance without distinguishing between a flu and a colic would lead to statistical averages devoid of decision-making value.

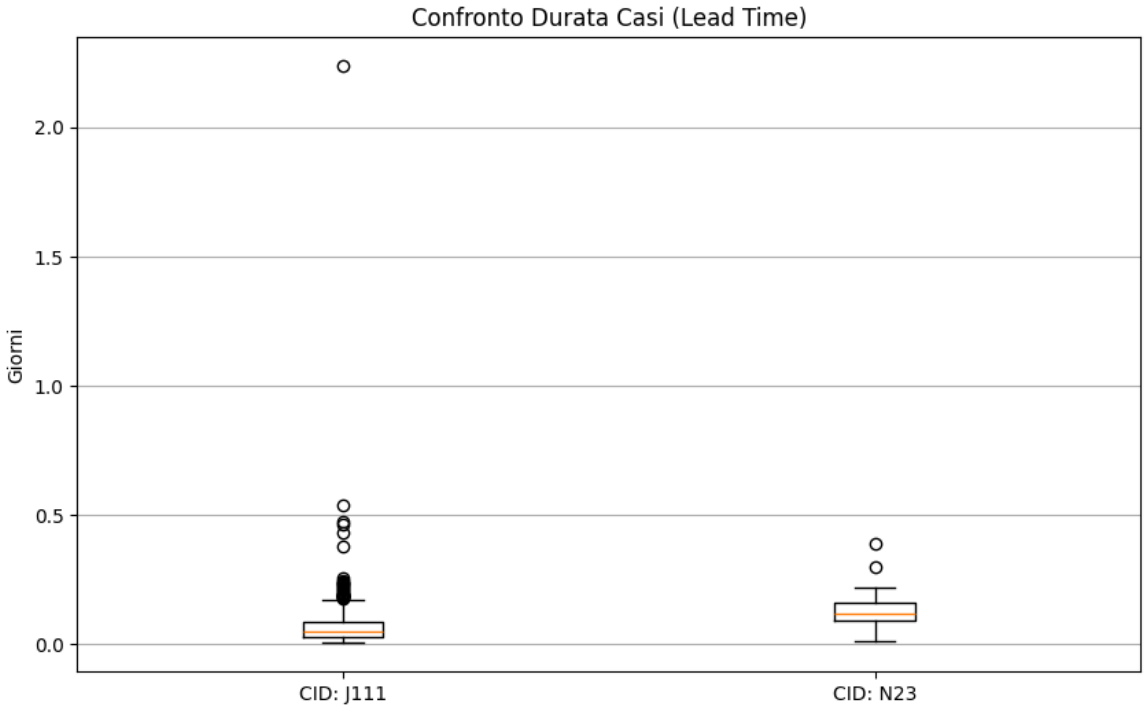
4.4 Performance Analysis: The Time Factor

On the two created segments, I calculated the **Case Duration (Lead Time)**, i.e., the total time elapsed from acceptance to discharge.

Statistical Results:

- **Influenza (J111):** Average duration of about **0.08 days** (approx. 2 hours). The variance is low.
- **Colic (N23):** Average duration of about **0.13 days** (approx. 3 hours and 10 minutes). The variance is much higher.

Clinical Interpretation:
The data show that patients suffering from colic remain in the hospital, on average, **60% longer** compared to patients with influenza. This figure is consistent with the clinical nature of the pathologies: colic often requires the administration of intravenous painkillers, an observation period to evaluate the effect of the drug, and sometimes complex diagnostic exams (ultrasounds/CT scans), whereas influenza is often resolved with a visit and a home prescription.



Click for full screen image

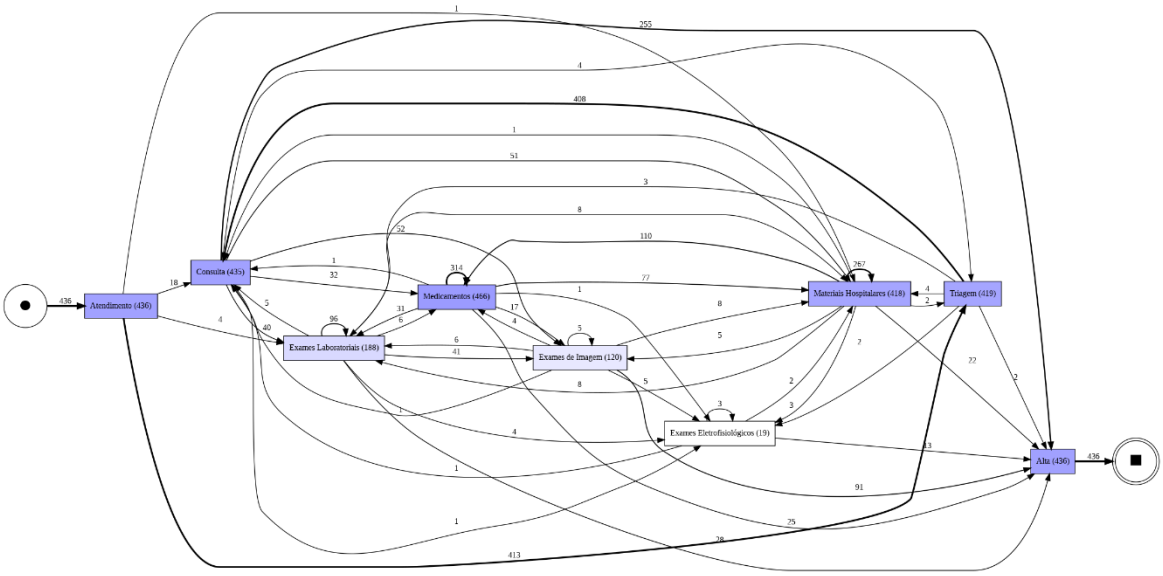
Comment on the graph: The Boxplot highlights not only the difference in medians (the central line) but also the different distribution. The "box" for Colic is wider and presents longer whiskers, indicating greater unpredictability in the treatment process compared to Influenza.

4.5 Process Discovery and Modeling

To visualize the flows, three modeling approaches supported by PM4PY were used.

A. Directly Follows Graph (DFG) - Frequency

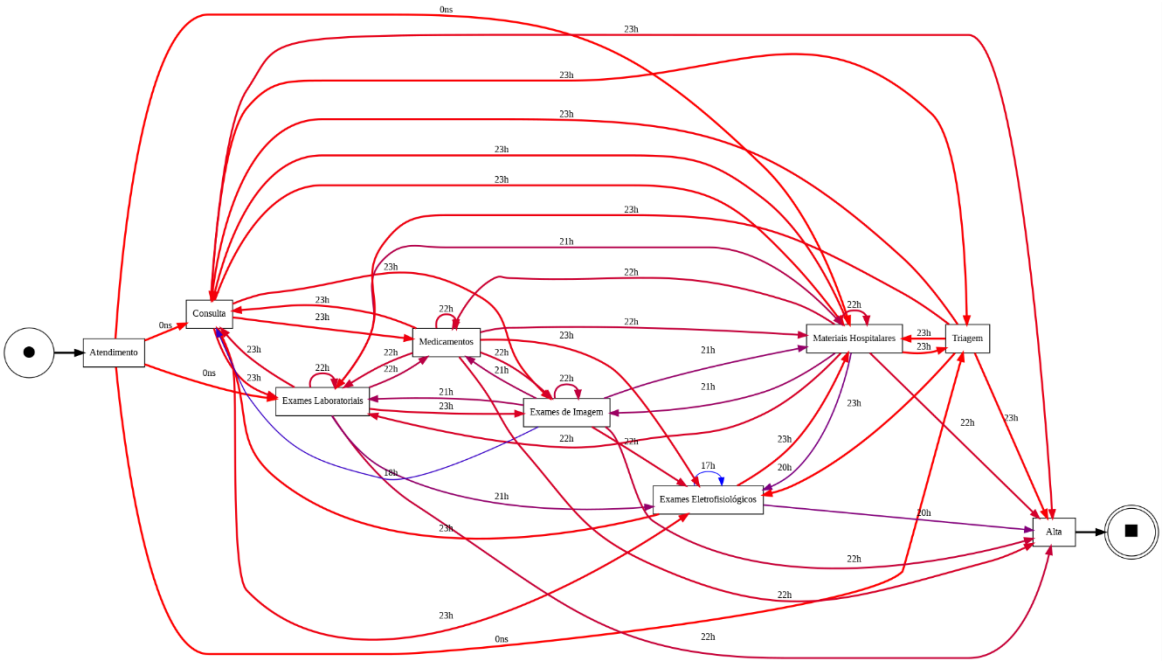
The graph clearly shows the main flow. However, **"loops"** (arrows going back) are visible between "Consulta" activities and exam activities ("Exames"). This indicates the **Rework** pattern: after an exam, the patient returns to the doctor for the reading of results. Although clinically necessary, from a process point of view, this represents a time burden.



Click for full screen image

B. Directly Follows Graph (DFG) - Performance

The performance-based DFG decorates the arcs with the average time elapsed. *Identification of the Bottleneck:* Visual analysis shows that the highest average times are not in the technical execution of exams, but in the waiting transitions. In particular, the time between Triage and First Visit (Consulta) is significant. This suggests that the availability of doctors is the limiting factor (scarce resource), not the administrative capacity of acceptance.



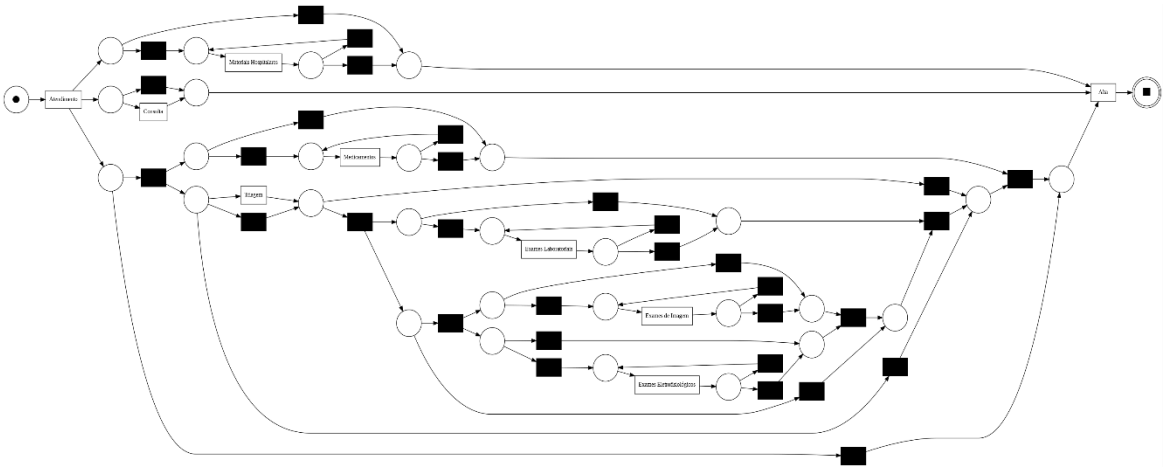
Click for full screen image

C. Petri Net (Inductive Miner)

Finally, the **Inductive Miner** algorithm was used to generate a Petri Net.

Unlike heuristic algorithms (which can create "spaghetti" models), the Inductive Miner guarantees the generation of a "sound" model (without deadlocks and with correct termination).

The resulting Petri Net clearly shows the logical structure: a sequential initial block, followed by an **exclusive choice (XOR split)** construct where the process branches towards different types of exams (Laboratory vs Imaging) or directly towards the exit, and finally a synchronization point (**AND join**) before discharge.



Click for full screen image

Conclusions

5.1 Synthesis of Empirical Evidence

The Process Mining project provided a transparent view ("White Box") based on data regarding the functioning of the Emergency Room, allowing for the validation of some hypotheses and the discovery of new ones.

1. **Healthy but Variable Process:** The facility efficiently manages the majority of cases (56% follow the ideal linear path), but suffers from a "long tail" of complex cases generating variability.
2. **Impact of Diagnosis:** It was statistically proven that the pathology (CID attribute) drastically influences the throughput time. The 60% difference in average times between Influenza and Colic demonstrates that treating all patients as a single flow is a managerial error.
3. **Bottleneck:** The main waiting time for the patient is located "upstream" of the actual clinical process, specifically in the wait for the doctor after triage.

5.2 Actionable Insights: Reorganization Proposals

Based on this evidence, and in response to the strategic objectives of Efficiency and Patient Centricity, the following improvement interventions are proposed:

1. **Implementation of a "Fast Track" for J111:**
   Given the high volume and low complexity of Influenza cases (J111), it is suggested to create a rapid path ("Fast Track"). This could involve a specialized nurse or a doctor dedicated exclusively to low-priority codes.
   Benefit: "Unloading" the main flow, reducing the load on resources that must manage complex cases like colic (N23) and cutting waiting times for everyone.
2. **Revision of the Triage-Consulta Process ("Advanced Triage"):**
   The bottleneck between Triage and Consulta suggests an imbalance of

resources. The use of advanced protocols allowing triage nurses to order standard exams (e.g., urine tests for suspected colic) *before* the formal medical visit could be evaluated.
Benefit: When the patient arrives at the doctor, the results could already be in processing, reducing the overall waiting time and speeding up the clinical decision.

3. **Administrative Automation:**
   Since the initial phase (Atendimento -> Triagem) is highly standardized, the introduction of digital kiosks for patient self-check-in with mild symptoms could reduce the administrative burden.

5.3 Original Contribution

The main original contribution of this work, beyond standard analysis, lies in the definition of a **domain-specific filtering heuristic**. Instead of applying blind statistical filters ("remove 5% of the rarest traces"), I implemented a logical filter based on the semantics of the healthcare process: a treatment is not finished if the patient is not discharged ("Alta"). This approach ensures that the calculated KPIs (such as average treatment time) are clinically valid and not polluted by system errors or dropouts, providing the hospital management with metrics that are much more reliable compared to a purely statistical approach.