NOVA IMS Information Management School

MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

Architecture for the application of Process Mining in a hospital setting

António Rodrigues Carvalho

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Architecture for the application of Process Mining in a hospital setting

by

António Rodrigues Carvalho

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Vítor Santos, PhD, Nova Information Management School

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, July 2024

ACKNOWLEDGEMENTS

First, I would like to express my deepest gratitude to Professor Vítor Santos who gave me indispensable guidance throughout the writing of this dissertation. Additionally, a special thanks to both interviewees for taking some of their time to provide helpful insights into the architecture.

To my family, my profound appreciation for all the support given to me over the last months.

ABSTRACT

The Portuguese population is ageing which puts more pressure on the public healthcare services. Therefore, it is required to improve the efficiency of these services in order to maintain a certain level of care quality to the patients. Nowadays, a hospital generates a significant amount of data patient's data, from vital signs to appointment scheduling, which offers an opportunity for revealing insights into the different processes. Process Mining, an emerging technology, provides the tools for discovering the different path patterns within the process as well as comparing them to pre-defined process models. This thesis proposes a holistic data architecture that can guide professionals into capitalizing hospitals' data by utilizing this technology in a hospital context. From the extraction of event data to the various Process Mining techniques, this architecture establishes a standard course of action based on the best practices in the field as well as the various types of analysis available. The architecture was also evaluated on its effectiveness by two experts of information systems in a healthcare context.

KEYWORDS

Process Mining; Healthcare; Healthcare Information Systems; Data Architecture; Health process

Sustainable Development Goals (SDG):

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

BPM	Business Process Management
CDM	Common Data Model
DSRP	Design Science Research Process
e.g.,	exempli gratia
EHR	Electronic Health Records
EMR	Electronic Medical Records
ER	Emergency Room
HIS	Hospital Information System
LIS	Laboratory Information System
OMOP	Observational Medical Outcomes Partnership
PM	Process Mining
RIS	Radiology Information System
SNOMED-CT	Systematized Nomenclature of Medicine – Clinical terms

1. INTRODUCTION

1.1. CONTEXT

Population ageing is considered to be one of the biggest challenges in the current days as health-related expenses increase significantly with age especially through the high hospitalization costs (Manole et al., 2023). In relation to Portugal, there is a public healthcare system (Sistema Nacional de Saúde-SNS) supported by the Portuguese state that ensures that every citizen has access to healthcare services, a right protected by law (Sousa, 2009).

The census from 2021 show an overview of the current situation of the Portuguese population through several indicators. The indicator showcasing the overall ageing of the Portuguese population is the average age which increased 3.1 years since 2011 to the age of 45.4 years in 2021. In certain areas of the country, the average age is even higher, with the regions of Alentejo and Center of Portugal registering 47.4 and 47.5, respectively. Additionally, there is an increasing percentage of population above the age of 65, while the active population and the population on the verge of entering the job market continue to decrease. The aging index, registered 182 people with the age of 65 or above for every 100 people between the ages of 0 and 14 whereas that value used to be 121 in 2011. Another perspective is the renewal rate of the active population which registers 76, in other words, for every 100 people leaving the job market only 76 are entering it. For comparison, that value used to be 94 in 2011 (Instituto Nacional de Estatística, 2022).

A study partly developed by Nova IMS reported that the Portuguese perceive a decrease in the overall quality of the services provided by the public healthcare system. One of the most pressing issues affecting the quality of the service are the excessive waiting times in health units and between appointments and the respective medical acts. Additionally, the overall access to medical services is perceived as limited by the users (AbbVie & Nova IMS, 2023).

Over the years, the conversion of data to the digital form has generated an increasing interest in exploring the different applications of Data Science in the healthcare industry(Ben Sassi & Yanes, 2023). This type of data is generated in large quantities and it comes from a wide range of sources, from treatment plans to patient's information. As such, there is a potential to gain insights from this data through the use of data science and big data analytics (Subrahmanya et al., 2022). The authors Kruse et al. (2016) summarize, through a systematic review, various advantages from the use of big data analytics in a healthcare setting. A significant amount of the analyzed papers focused on boosting the overall efficiency of care and the early detection of diseases which support the monitoring of the patient's medical condition(Kruse et al., 2016).

1.2. MOTIVATION

Various hospitals from the Portuguese public healthcare system are facing increasing operational expenses (Conselho das Finanças Públicas, 2024). Given this, according to Toussaint & Berry (2013), enhancing the operational efficiency of hospitals is the key to cut down operational expenses while maintaining or even improving the care given to the patients. They also give a few examples of the efforts implemented by hospitals in the United States to drive up the efficiency. One of them is the enactment of Lean principles, which is a shift of the organizational culture characterized by the constant seeking of improvements to the operations (MHPC, 2014). Some of the mentioned impacts of Lean in a healthcare setting include speedier operating room turnover and response time for urgent cases(MHPC, 2014).

An emerging field focused solely on supporting the improvement of processes in an organization is Process Mining (W. M. P. van der Aalst, 2011) . It is the link between traditional methods of process optimization (e.g., simulation, BPM techniques) and advanced analytics (e.g., Machine learning, Data Mining)(Mans et al., 2015). The application of Process Mining in health is being increasingly studied as the number of related publications has been growing since 2013. The focus of the analysis extends to a wide variety of processes including clinical-related processes (e.g., medical treatment) and organizational processes (e.g., billing) (De Roock & Martin, 2022).

The rising interest of PM by organizations is accompanied by a growing demand for analysts specialized in the field. Nonetheless, entry barriers still persist as there is a lack of effort to support knowledge transfer (Zimmermann et al., 2023). Therefore, the main motivation behind this dissertation is to provide professionals the required support for them to learn more efficiently the standard set of actions and best practices when applying PM in a hospital setting in order to capitalize on the data generated by the hospital as much as possible. In turn, the resulting insights will contribute to the rise of the processes' efficiency and reduce operational costs along with the increase of the quality of care given (Munoz-Gama et al., 2022).

1.3. OBJECTIVE

The objective of this dissertation is to propose an architecture for the application of Process Mining in a hospital setting to spread its use across several health organizations inside the Portuguese public healthcare system and in turn contribute to the sustainability of the latter. In order to achieve this goal, the following intermediate objectives were defined:

- Creation of a comprehensive study on the use of Process Mining in the healthcare industry.
- Build of an architecture for the use of process mining in a hospital setting
- Evaluation of the architecture

2. LITERATURE REVIEW

2.1. OVERVIEW OF A HOSPITAL

2.1.1. Hospital Organizational Structure

Organizational charts are key enablers to comprehend the operations inside an institution(Ostroff, 1999). For that reason, it was used as an example the organizational structure of a clinical establishment Centro Hospitalar Lisboa Norte (CHLN), which is a fusion between Hospital Santa Maria and Hospital Pulido Valente, both situated in Lisbon, Portugal.

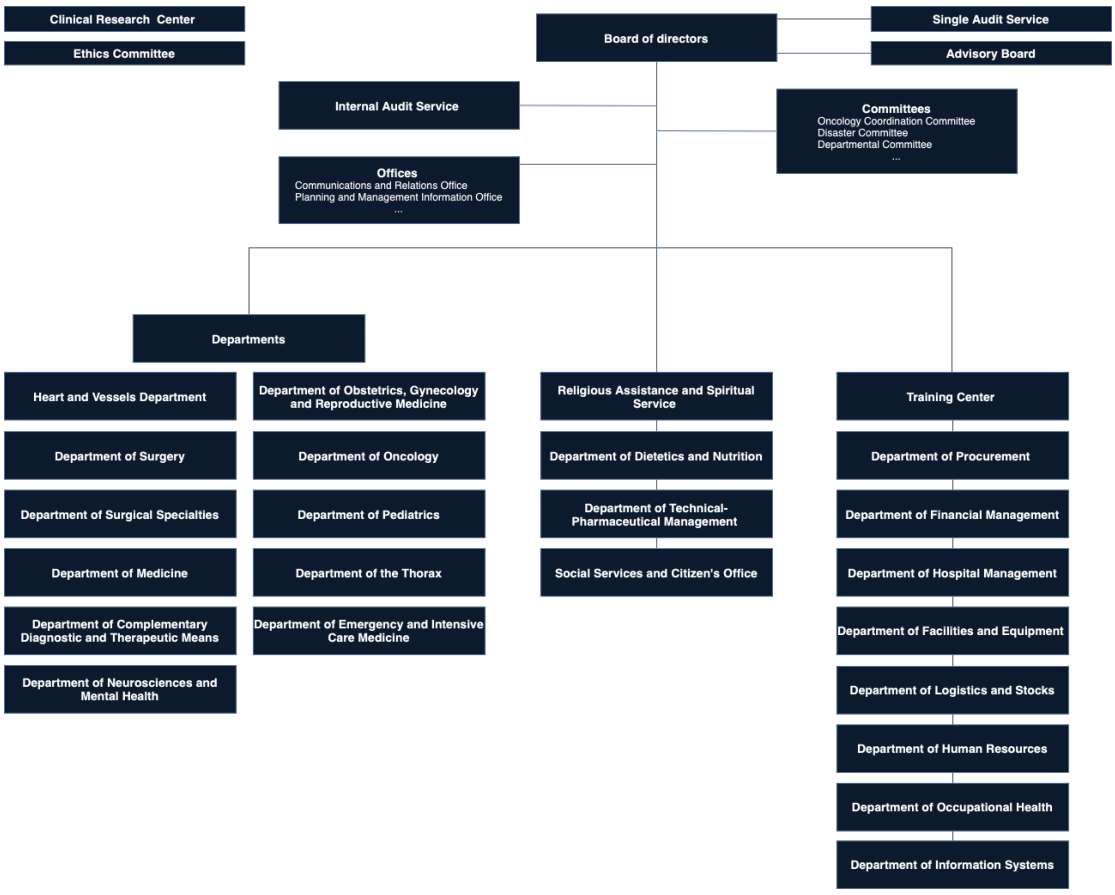

Figure 1 - CHLN's Organizational Chart adapted from CHLN (2016)

According to the figure 1, the structure of CHLN is divided into 3 main areas: Medical work, Support for the provision of care and General Support and Logistics. The medical work is organized in a "matrix structure based on departments, services and functional units" as a way to foment the joint work between the different teams to tackle each pathology. In other words, this structure aims at creating a new work dynamic that centers around the patient's well-being. From figure 1, the division of the several departments is made through several through the type of pathologies (e.g., Department of Oncology). The other two areas are more focused on supporting the medical actions but through a logistical and managerial

perspective. There is a wide variety of services required for the functioning of the institution, from services of procurement to financial management(CHLN, 2016).

2.1.2. Healthcare Processes

Lillrank & Liukko (2004) established a general notion of a process in a healthcare setting. According to the authors, a process is constituted by a set of human resources (e.g., physician, nurses) and material resources (e.g., medical supplies) which have the patient's medical condition as an input and through an assessment (e.g., diagnosis) followed by the application of an algorithm (e.g., medical treatment) it outputs again the patient's medical condition.

Mans et al. (2015) introduces a division of the healthcare processes centered about patients' care which according to Munoz-Gama et al. (2022) is ultimately an important indicator of care quality provided by the hospital.

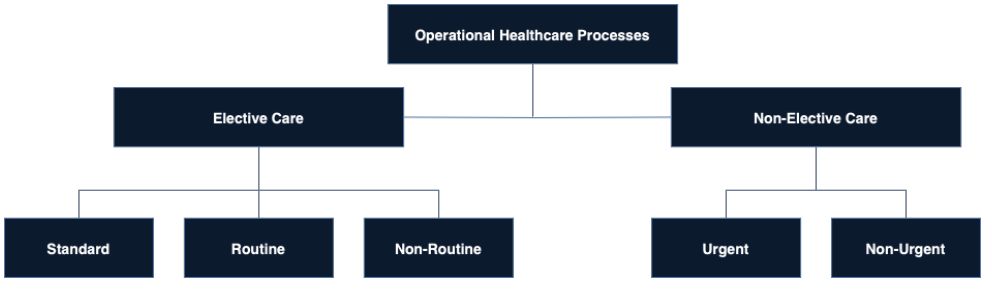

Figure 2 - Division of operational healthcare processes adapted from Mans et al. (2015)

Figure 2 showcases an overview of the different types of process which at the first level is divided based on the urgency level of care required by the patient's medical condition. In elective care, the health condition of the patient is stable enough that it allows the deferral of the treatment for a days or weeks. In opposition, non-elective care covers all cases where it is required a fast intervention. Within elective care, the processes differ according to their variability level, which is based on the division of healthcare processes proposed by Lillrank & Liukko (2004). For standard processes, the process does not vary along the patients as the treatment and the respective result are considered a given. However, for routine processes the treatment can change sporadically for a patient based on certain circumstances. Finally, for non-routine processes, the medical treatment for each patient is different due to their sensitive medical condition. On the other side, within non-elective care, the division is again based on the required response time. Urgency care refers to the care that should be immediate, meanwhile non urgent care can be postponed for a certain amount of time.

2.2. PROCESS MINING

The next chapter introduces the overall concept of Process Mining and its technicalities. Additionally, it gives a practical perspective as it presents the context under which Process Mining can be used and the limitations it faces.

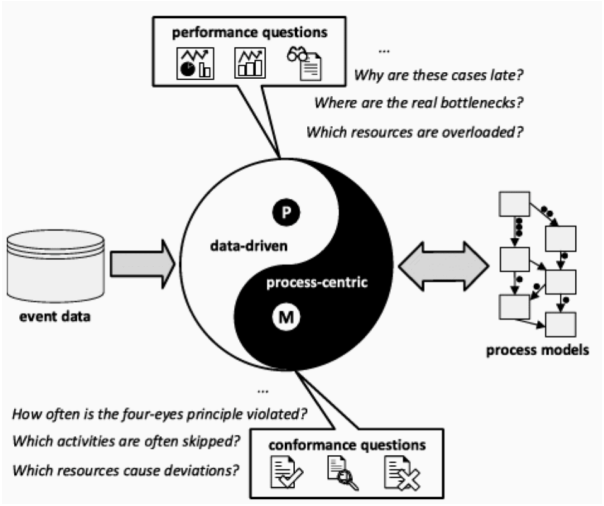

Figure 3 - Overview of Process Mining (W. van der Aalst, 2016)

According to van der Aalst (2016), Process Mining (PM) is considered to be the connection between Data Science and Process Science. Figure 3 provides an overview of Process Mining, from the source to the type of questions it tries to answer. The starting point of PM is event data, which when analyzed in both a data and process centric manner, it allows to uncover insights into the process in terms of its efficiency (e.g., bottlenecks) and structure (e.g., activity sequence).

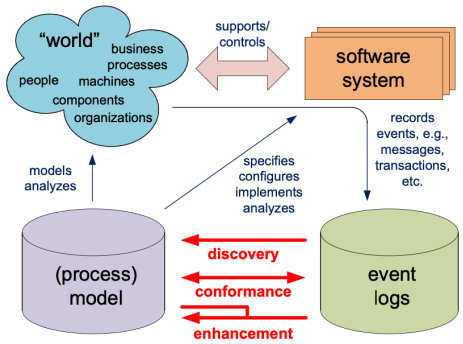

Figure 4 - High level architecture of Process Mining (W. van der Aalst, Adriansyah, de Medeiros, et al., 2012)

Figure 4, in comparison with figure 3, offers a more in-depth view in how PM operates. Event data, referred previously, should be in the form of event log and can be originated from a variety of sources such as event records or transactions. Afterwards, event logs can be used as inputs for the different PM techniques (Process Discovery, Conformance Checking, Process Enhancement) depending on the end goal. The most well-known technique is Process Discovery which, based on the event log, outputs a process model. On the other hand, Conformance Checking compares process models with the event logs for the purpose finding any inconsistency between expectation and reality. At last, Process Enhancement improves the process model, previously designed, with the added information provided by the event log(W. van der Aalst, Adriansyah, de Medeiros, et al., 2012).

2.2.1. Event Log

In order to explain the functioning of an event log, Mans et al. (2015) details the structure of the event log through the relationship between its several components (e.g., case, event) in terms of cardinality alongside the different levels of magnitude where each component is situated.

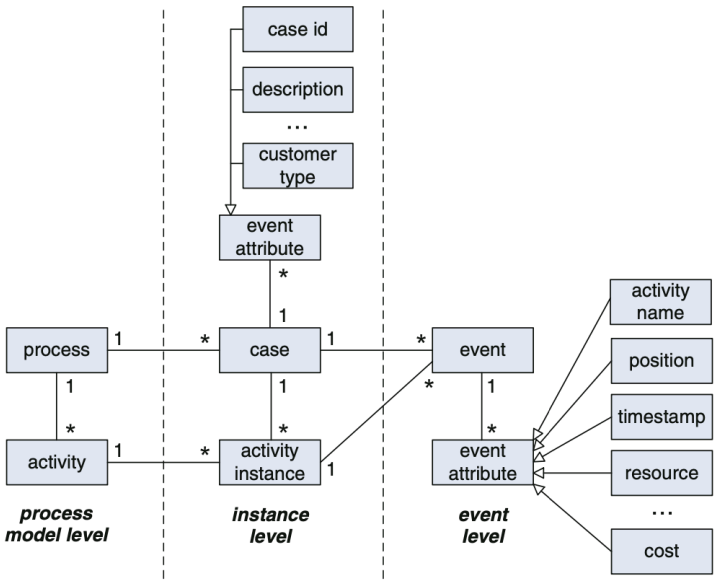

Figure 5 - Required information in the event log (Mans et al., 2015)

Figure 5 describes the several layers of the hierarchy established in the event log which starts at the event level, solely composed of events. Then, at the instance level, each event belongs to a single activity instance which on its own is associated to a particular case. In turn, all cases are related to the one single process contained in the event log. Additionally, the event log can be enriched with the assignment of attributes (e.g., resource) to each event or case. At last but not least, the authors also stress the importance of having an attribute which can represent the time ordering of the events in process either through an index or timestamp(Mans et al., 2015).

W. M. P. van der Aalst (2011) simulates the extraction process of event logs from a database through flattening the data model. This method is under the assumption that activities correspond to a change of the status of the case. As such, the only rule mentioned is that each event must be associated with a case.

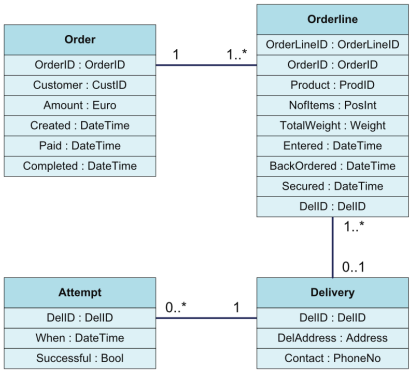

Figure 6 - Database diagram (W. M. P. van der Aalst, 2011)

Case id	Activity	Timestamp	Other attributes
91245	Create order	28-11-2011:08.12	Customer: John, Amount: 100
91245	Enter order line	28-11-2011:08.13	OrderLineID: 112345, Product: iPhone 4G, NofItems: 1, TotalWeight: 0.250, DellID: 882345
91245	Enter order line	28-11-2011:08.14	OrderLineID: 112346, Product: iPod nano, NofItems: 2, TotalWeight: 0.300, DellID: 882346
91245	Enter order line	28-11-2011:08.15	OrderLineID: 112347, Product: iPod classic, NofItems: 1, TotalWeight: 0.200, DellID: 882345
91245	Secure order line	28-11-2011:08.55	OrderLineID: 112345, Product: iPhone 4G, NofItems: 1, TotalWeight: 0.250, DellID: 882345
91245	Create backorder	28-11-2011:08.55	OrderLineID: 112346, Product: iPod nano,

Figure 7 - Snippet of an Event log adapted from W. M. P. van der Aalst (2011)

The demonstration of the method in question was done with the support of the database represented in figure 6. First and foremost, it was arbitrarily chosen each order as the case and as it is connected to the rest of the tables, it is possible to connect possible events from those tables ('Orderline','Delivery' and 'Attempt') to each order. The extraction of the events is done through columns containing dates, for instance, the value in column 'Created ' from table Order, shown in figure 6, will correspond to the event 'Create order', shown in figure 7. The event log from figure 7 shows several events such as 'Create order' and 'Enter order line' from tables 'Order' and 'Orderline',respectively, and in which all events are related to one particular order, order number 91245 (W. M. P. van der Aalst, 2011).

2.2.2. Process Discovery

Out of the several PM techniques, Process discovery is the one that stands the most for its significance (W. van der Aalst, Adriansyah, de Medeiros, et al., 2012). The main objective is to build a process model that can portray the behavior of the process (Mans et al., 2015) . Therefore, it should receive an input in the form of an event log and output a Process model (van Eck et al., 2015), which can be portrayed in several notations (e.g., BPMN, Petri Net)(Mans et al., 2015). Important to address that most notations allow the showcase of other perspectives other than simply the ordering of activities (control-flow) such as resource perspective through showing the entity (e.g., person, department) responsible for the execution of the activities.

2.2.3. Process Conformance

Technique that enables the comparison between an arbitrary Process model and what is the actual process, which is recorded in the even log (W. van der Aalst, Adriansyah, & van Dongen, 2012). More specifically, it checks for deviations between the process model and the event log in terms of the trace of activities (Mans et al., 2015). W. M. P. van der Aalst (2011) introduces the two types of analysis regarding deviations:

- Global Conformance Measures: Global conformance between the process model and the event log, for instance 'Only 50% of the cases in the event log can be reproduced exactly as in the process model'.
- Local diagnostics: Narrowing the analysis into specific activities which are shown to be the source of deviations, for example 'Activity X was executed more times than the process model allowed'.

2.2.4. Process Enhancement

This process consists of enacting changes to the Process model defined a-priori to better fit reality through the information given by the event logs (W. van der Aalst, Adriansyah, & van Dongen, 2012). The process model can either be repaired in terms of adjusting the already existing information or can be extended with new information provided by the event log(Mans et al., 2015). This extension can be made through adding a new perspective (e.g., resource, time) to the process model given the attributes available in the event log, for instance it is possible to add a time perspective into the process model in case there are associated timestamps to the events in the log (W. van der Aalst, Adriansyah, & van Dongen, 2012).

2.2.5. Process Mining based Methodology

In a systematic review regarding the application of PM in healthcare, Rojas et al. (2016), it was found that 12% of the analyzed articles propose a PM based methodology containing a set of guiding principles to help implement PM. Methodologies such as L∗ life-cycle model or PM² contributed to the increase usage of PM by researchers from other areas (Munoz-Gama et al., 2022).

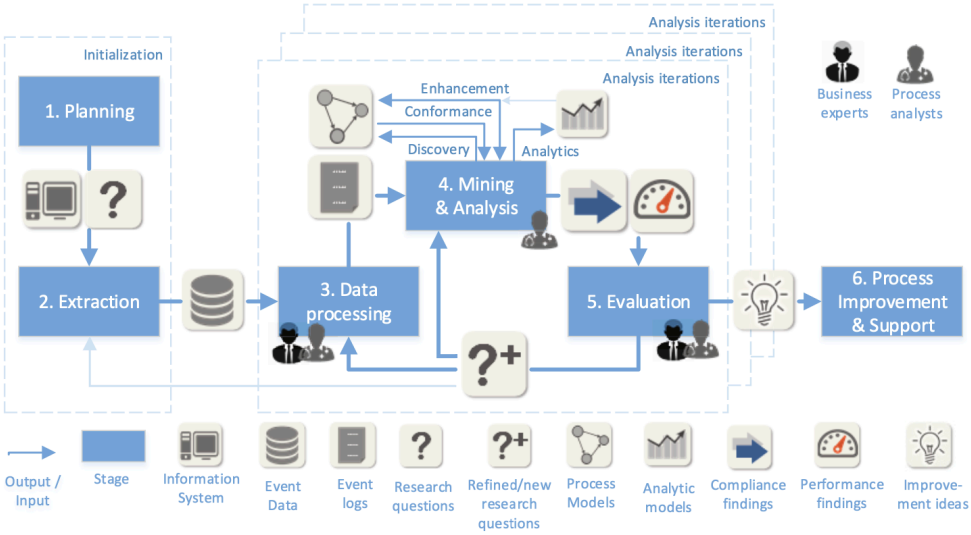

Figure 8 - Overview of PM² adapted from van Eck et al. (2015)

Figure 8 shows an overview of the methodology of PM² which ultimately attempts to enhance the performance of the process and the respective compliance towards regulations or guidelines mainly through PM techniques. The methodology, according to van Eck et al. (2015) is composed by 3 phases:

- Phase 1: This phase also known as 'Initialization' is composed by the Planning and Extraction stages. Essentially, during this phase it is defined the focus of the project through research questions and the extract of event data.
- Phase 2: Event data once extracted goes through an analysis phase which is composed by 3 stages: 'Data Processing', 'Mining & Analysis' and 'Evaluation'. These stages cover the process of turning event data into an event log ('Data Processing'), analyzing the event log through PM techniques ('Mining & Analysis') and then deciding whether the resulting insights can answer the research questions ('Evaluation'). If that is not the case, then researchers have the option to iterate over the latter stages until they reach the satisfying answers. Over each iteration, changes can be made within each stage, for instance, by creating a different view for the event log during 'Data Processing'.
- Phase 3: The last phase, which only contains stage 'Process Improvement & Support', aims at capitalizing the insights from the 'Evaluation' stage and alter the execution of the process accordingly.

2.3. Process Mining for Healthcare

The quality level of the services offered by the hospital is directly proportional to how efficient it is when executing the respective processes(Rojas et al., 2016), whether organizational or clinical (Lenz & Reichert, 2007). Process Mining when applied in this context gives the possibility to make a deep diagnosis to these same processes, which range from finding patients' clinical path to bottlenecks in the services(*Cancer Diagnostic Delay Reduction*,

n.d.).With this said, over the last 10 years there has been a steady increase of publications with a focus on Process Mining in health-related services, which goes to show the potential of this area (De Roock & Martin, 2022). Given this, on this chapter it will be discussed the latest status of process mining in the healthcare industry. This discussion will be done through a systematic revision of literature. This type of study has as its foundation on the analysis of all articles relevant for the topic in question in order to be the most rigorous and unbiased possible (Caldwell & Bennett, 2020). For this purpose, it was chosen PRISMA (Preferred Reporting Items for Systematic Reviews and Metanalysis) method, which is characterized by collecting articles from several scientific databases (e.g., SCOPUS), filter them according to certain criteria and finally discuss them (Liberati et al., 2009).

Table 1 - Questions to be answered by PRISMA

	Questions
RQ1	What is the current state of the art?
RQ2	What are the processes eligible for Process Mining?
RQ3	What is the standard pre-processing treatment to the data?
RQ4	What are the most used methodologies?
RQ5	What are the most used data analysis techniques?

2.3.1. Systematic Literature Review Execution

In the first phase of selection (Identification), it is selected all the articles related with the topic. In this case, the articles were source from 3 different scientific databases to cover a wider range of articles. It is also important, due to the constant flux new articles, to stress out that the search was made in December 2023.

Table 2 - URL of the scientific databases

Databases	**URL**
Scopus	https://www.scopus.com/
IEEE Xplore	https://ieeexplore.ieee.org/
Web of Science	https://www.webofscience.com

The search inside these databases was done with the support of several key words related to Process Mining and Healthcare concepts discussed in the this chapter previous sections . In total, it was found 303 documents.

Table 3 - Search Strings

Databases	KeyWords
Scopus	"Process Mining" OR "Process Discovery" OR "Conformance Checking" OR "Performance Mining" AND ("Health" OR "Healthcare" OR "Hospital")
IEEE Xplore	("Document Title":Process OR "Document Title":Process Mining OR "Document Title":Process Discovery OR "Document Title":Conformance Checking OR "Document Title":Performance Mining) AND ("Document Title":Health OR "Document Title":Healthcare OR "Document Title":Hospital)
Web of Science	("Process Mining" OR "Process Discovery" OR "Conformance Checking" OR "Performance Mining") AND ("Health" OR "Healthcare" OR "Hospital")

In the screening part, there was the adding of more filters. All the articles that do not present one of the characteristics were excluded. The first filter was excluding all articles published before 2020 and as such it was filtered out 163 papers. The second one was solely including academic papers in the form of articles, excluding other 76 papers. Another exclusion filter was for the articles to be all in English, in which all of them were as such none were excluded. At this stage, the articles from the several scientific were merged, which enabled us to remove duplicated papers accounting for the removal of 28 articles. For the title/abstract/full text screening, any article that was not related with the topic of Process mining in a healthcare context was dismissed, it was dismissed 9, 4 and 0, respectively. All in all, 21 articles were finally selected for further analysis, as it is shown below in figure 9.

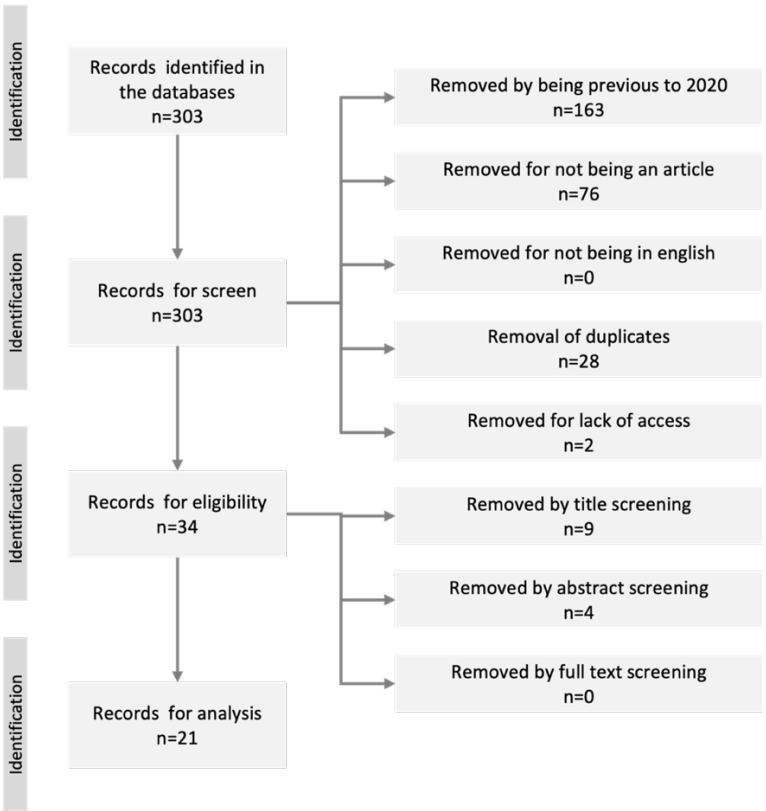

Figure 9 - Diagram of the selection of articles

Table 4 - Final selection of articles

ID_PAPER	Title	Reference	Application
P1	Exploring the potential of OMOP common data model for process mining in healthcare	(Park et al., 2023)	Extraction of event logs from OMOP based databases and the respective PM analysis
P2	Applying process mining in health technology assessment	(Dallagassa et al., 2022)	Application of PM for evaluating two different health related technologies
P3	Process mining framework with time perspective for understanding acute care: a case study of AIS in hospitals	(Pang et al., 2021)	Development of a PM framework with a focus on time perspective
P4	Evaluation of patient transport service in hospitals using process mining methods: Patients' perspective	(Kropp et al., 2023)	Application of PM for the analysis of patient transport service process
P5	Conformance Checking: Workflow of Hospitals and Workflow of Open-Source EMRs	(Asare et al., 2020)	Conformance Checking between event logs provided by an open source EMR (OpenEMR) and domain based process model
P6	Towards the use of standardized terms in clinical case studies for process mining in healthcare	(Helm et al., 2020)	Investigation into the potentials of employing standardized clinical codes
P7	Process mining project methodology in healthcare: a case study in a tertiary hospital	(Pereira et al., 2020)	Proposal of enhanced version of THE methodology PM² adapted to the health context
P8	Mapping the patient's journey in healthcare through process mining	(Arias et al., 2020)	Application of PM techniques for the assessment of the customer journey
P9	Privacy-preserving process mining in healthcare	(Pika et al., 2020)	Comparison between different data transformation techniques for data privacy
P10	Supporting Governance in Healthcare Through Process Mining: A Case Study	(Agostinelli et al., 2020)	PM analysis through 3 different perspectives
P11	Process Mining approach for discovering and analyzing the healthcare processes in Python	(Rashed et al., 2023)	Exploration of the potentials of Python as a PM tool
P12	Improving the In-Hospital Mortality Prediction of Diabetes ICU Patients Using a Process Mining/Deep Learning Architecture	(Theis et al., 2022)	Use of PM as a supporting tool for predictive analysis

P13	Dynamic models supporting personalized chronic disease management through healthcare sensors with interactive process mining	(Valero-Ramon et al., 2020)	Construction of Dynamic risk models based on health sensors data with PM as support
P14	Process mining as support to simulation modeling: A hospital-based case study	(Tamburis & Esposito, 2020)	Development of a DES model with the input of PM
P15	How Can Interactive Process Discovery Address Data Quality Issues in Real Business Settings? Evidence from a Case Study in Healthcare	(Benevento et al., 2022)	Investigation of a new PM technique named Iterative Process Discovery (IPD)
P16	Process mining for healthcare: Characteristics and challenges	(Munoz-Gama et al., 2022b)	Summary of specificities of PM in a healthcare setting and the respective challenges
P17	Performance Analysis and Activity Deviation Discovery in Event Log Using Process Mining Tool for Hospital System	(Sundari & Nayak, 2022)	Use of PM related algorithms for the discovery of deviation flows
P18	Extending a Data Management Maturity Model for Process Mining in Healthcare	(Erhard et al., 2023)	Inclusion of a data quality dimension to a data maturity model (RAMM)
P19	Recommendations for enhancing the usability and understandability of process mining in healthcare	(Martin et al., 2020)	Formulation of recommendations for the development of PM projects in a healthcare setting
P20	Process mining and lean six sigma: a novel approach to analyze the supply chain quality of a hospital	(Ramires & Sampaio, 2022)	Exploration of the cross between PM and Lean Six Sigma tools to analyze a supply chain related process
P21	Applying process mining and semantic reasoning for process model customisation in healthcare	(Pereira Detro et al., 2020)	Construction of process models with the supply of PM and domain knowledge through Ontologies

2.3.2. Discussion

RQ1: What is the current state of the art?

Process Mining in combination with other areas is a topic that is discussed in several articles, from integration with deep learning, from P12 (Theis et al., 2022) to health sensors from P13 (Valero-Ramon et al., 2020). Valero-Ramon et al. (2020) demonstrates the use of process mining with real time data provided by health sensors. The consequent insights regarding the patient's medical condition and the overall health progression data are then used for the construction of dynamic risk models for pathologies including diabetes and hypertension. The combination of Process Mining and Deep Learning present in P12 (Theis et al., 2022) is the prediction of mortality rate of patients in the ICU (Intensive Care Unit) that suffer from

diabetes. More specifically, historical data is used to create a process model with the support of a technique, Decay Replay Mining, which distributes weight to the events in accordance with their recency. Afterwards, the process model and recent patient's data are fed into an algorithm which predicts the risk of mortality. Another example, from P14 (Tamburis & Esposito, 2020), exemplifies the creation of Discrete Event Simulation (DES) model with the support of Process Mining, which tries to simulate cases of the process of patients in the ophthalmology ward. The creation of the DES model is based on relevant information from the patterns in the process, encountered by process mining. The model, based on the process's behavior in the past, can simulate the process under different scenarios which can give valuable insights into its overall performance.

Another topic of discussion present in papers P3, P6 and P19 is the standardization of clinical terms in the event log. Article P6 (Helm et al., 2020) focuses on the importance of standardizing clinicals terms for each activity in the event log during the pre-processing stage for increased understandability of the process as well as comparability across other research papers. The standard clinical terminologies used in the paper were ICD 10 and SNOMED CT and both were evaluated positively in terms of being able to respectively assign codes for all types of clinical environments and diagnosis. Article P3 (Pang et al., 2021) also mentions an example of other terminologies such as CCHI (Chinese Classification of Health Interventions) or MCCSI (Medicine Classification and Codes for Social Insurance). The lack of standard terminology for concepts of process mining (e.g., activity log and event log) is also mentioned in article P19 (Martin et al., 2020) which is an issue across research papers.

RQ2: What are the processes eligible for Process Mining?

Several articles including P17 (Sundari & Nayak, 2022) and P4 (Kropp et al., 2023) mention a division of the hospital processes into two perspectives, medical and organizational. The processes of the medical type, as the name suggests, are directly related to the patient's health treatment. Meanwhile, organizational processes include activities that enable the functioning of the hospital itself (e.g., billing process). For instance, P4 (Kropp et al., 2023) studies a process that involves the transport service of patients inside the hospital from the request stage to the transportation of the patient.

Regarding medical processes, P1(Park et al., 2023) proposes a further division of these processes based on the type of visit to the hospital: Inpatient, Outpatient, Emergency Room and Patient Journey. The exact difference between Inpatient and Outpatient visits has not been standardized but the most prominent notion is that visits that outlast a day (LOS>1) are considered inpatient whereas for visits with less duration (LOS<1) are considered outpatient(Bovonratwet et al., 2017). Emergency room (ER) processes represent cases of patients that require immediate care by going to the ER. Finally, patient journey process is different from the rest of the processes as it includes the entire medical history of the patients over a span of time and as such it gathers the other 3 types of visits.

Case ID	Event ID	Activity	Timestamp	Originator	Department	Sex	Ages
C10001	E1001	Physical examination	2021-08-01 09:03:00	O10001	Nursing dept.	Female	30
	E1002	Consultation	2021-08-01 10:01:00	O10002	Hematology oncology dept.	Female	30
	E1003	Lab test	2021-08-01 12:30:00	O10003	Internal Medicine dept.	Female	30
	E1004	Drug	2021-08-01 12:35:00	O10004	Pharmaceutical care dept.	Female	30
C10002	E1005	Consultation	2021-08-02 10:01:00	O10005	Plastic Surgery dept.	Male	50
…	…	…	…	…	…	…	…

Figure 10 - Sample of an event log of an outpatient process adapted from Park et al., 2023

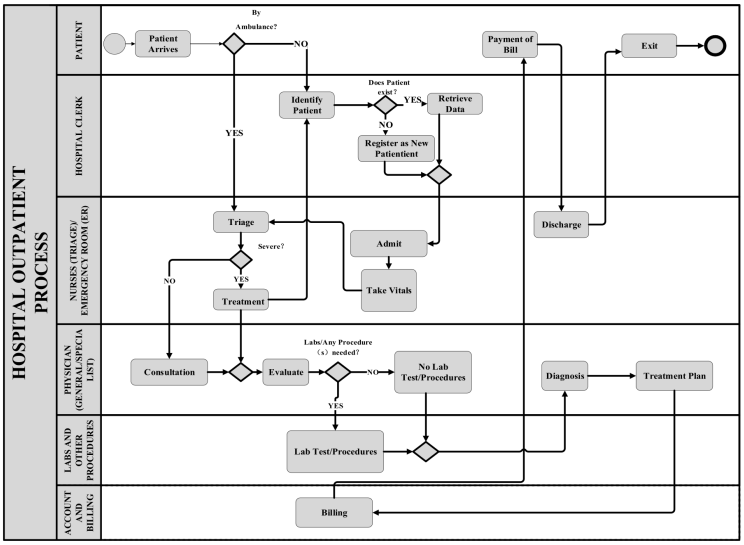

Figure 11 - Example of an outpatient process (Asare et al., 2020)

According to the American Hospital Association, Outpatient visits are defined by visits from patients who are not lodged in the hospital (CDC, 2023), as shown in figure 11, taken from P5 (Asare et al., 2020), which does not contain any information regarding hospitalization. P1(Park et al., 2023) refers a tendency from past research for grouping activities into an operational level for both clinical activities (e.g., consultation) and non-clinical activities (e.g., Payment), which can also be seen in the figure 11 and in the event log from figure 10. According to P1 (Park et al., 2023), the analysis for this type of visit is mainly characterized by process discovery followed by performance indicators such as waiting times. Additionally, in P10 (Agostinelli et al., 2020), the focus is on the organizational perspective of Radiology department and its sub-departments. For instance, it is used social networks to find the relationship between the sub-departments in terms of how patients utilize these resources.

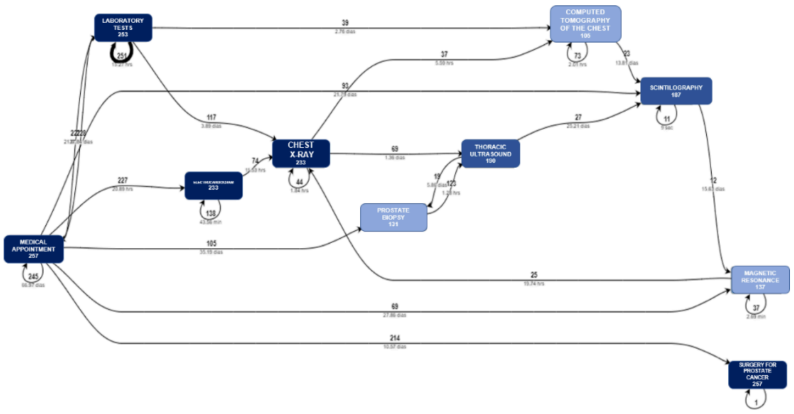

Figure 12 - Example of a Pre-Surgery care process (Dallagassa et al., 2022)

Case ID	Event ID	Activity	Timestamp	Originator	Length of stay (LOS)	Sex	Ages
C10001	E1001	Fentanyl 0.1 MG	2021-08-01 09:03:00	O10001	3	Female	40
	E1002	Midazolam 5 MG/ML	2021-08-01 10:01:00	O10002	3	Female	40
	E1003	Segmented neutrophyls/100 leukocytes in blood by automated count	2021-08-01 12:30:00	O10003	3	Female	40
C10002	E1005	Fentanyl 0.1 MG	2021-08-02 10:01:00	O10005	4	Male	40
...	...	...	...	...	...	...	...

Figure 13 - Event log from an inpatient process adapted from Park et al., 2023

For inpatient processes, according to P1 (Park et al., 2023), the tendency of research is instead in the clinical pathway of the patients, more particularly, the set of clinical actions which make up the entirety of the treatment of the patient as demonstrated in the figure 12, which includes activities such as 'Chest X-RAY' or 'Magnetic Resonance', part of the Pre-Surgery care process. Figure 13, taken from P1 (Park et al., 2023), shows that the granularity in the event log can also be higher as it shows the exact medication given to the patients or the count of specific blood cells during blood tests which may be important to analyze. Other examples of inpatient process include pre and post-surgical care processes from P2 (Dallagassa et al., 2022) or the clinical path of patients with the same pathology from P8(Arias et al., 2020).

The high variety of patient processes come also with variety in the respective analysis. For instance, in P8 (Arias et al., 2020) it was applied variant analysis in conjunction with Length of Stay and other attributes (e.g., age, gender), where for example it was found that older patients required more clinical activities such as medical images or procedures. In the case of P2 (Dallagassa et al., 2022), it is used KPIs for a comparison between two medical treatments, more specifically operational-related indicators (e.g., average cost, length of stay) as well as clinical-related indicators (e.g., death rate, relapsed rate). P3(Pang et al., 2021) also compares

two different treatments for the same pathology, Acute ischemic stroke (AIS). Besides analyzing the control-flow it was also used the time related KPIs specific to AIS, Imaging to Needle time (INT) and Door to Needle time (DNT), which then were used to compare the hospital's performance with the recommended values issued by clinical guidelines.

Case ID	Event ID	Activity	Timestamp	Originator	Department	Sex	Ages
C10001	E1001	Visit from Home	2021-08-01 09:03:00	O10001	ER dept.	Female	30
	...	...	...	...	...	...	...
	E1003	Cooperative care	2021-08-01 12:30:00	O10003	Pediatric dept.	Female	30
	...	...	...	...	...	...	...
	E1009	Discharge to home	2021-08-01 12:30:00	O10003	ER dept.	Female	30
C10002	E1011	Patient transfer from hospital to hospital	2021-08-02 10:01:00	O10005	Plastic Surgery dept.	Male	50
...	...	...	...	...	...	...	...

Figure 14 - Event log from an ER process (adapted from Park et al., 2023)

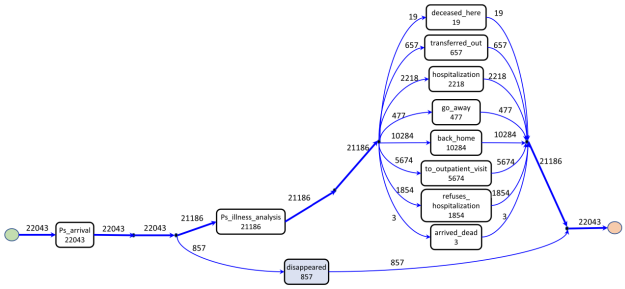

Figure 15 - Example of an ER process (Agostinelli et al., 2020)

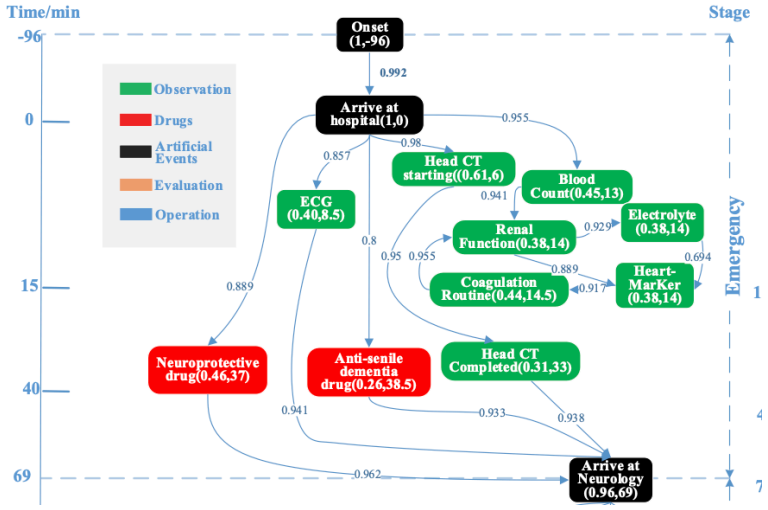

Figure 16 - Example of an ER process adapted from Pang et al., 2021

According to figure 15 and figure 16, ER process is generally characterized by the patient's arrival to the hospital and the consequent transfer to other medical Units (e.g., Neurology) or other events such as being sent home or transfered to another hospital. Given this, most medical activities seem to be either simple medical activities such as observation or drug administration or transfers between places as it can be shown in figure 14. On article P10 (Agostinelli et al., 2020), the analysis centered around operational point of view regarding

patients' waiting time specially the relationship between patients that left the facilities before the medical observation and the respective waiting time through a frequency distribution. One thing to note is the fact that ER process can be integrated in the previous two processes as figures 16 and 11 show the integration with inpatient and outpatient process, respectively. Therefore, depending on the intended analysis, ER process can either be analyzed separately or in conjunction with the other processes. Likewise, the level of granularity can also vary with the decision to integrate or not with other processes. Figure 16 shows a snippet of an ER process in the overall inpatient process and it is more detailed than figure 15 as it shows the exact examinations and medications administered to the patients whereas figure 15 generalizes all these activities into 'PS_Illness_analysis'.

Case ID	Event ID	Activity	Process type	Timestamp	Originator	Department	Sex	Ages
C10001	E1001	Outpatient visit	9202	2021-08-01 09:03:00	O10001	ER dept.	Female	30
	E1003	Inpatient visit	32036	2021-08-01 12:30:00	O10003	Pediatric dept.	Female	30
	E1009	Emergency room visit	9201	2021-08-01 12:30:00	O10003	ER dept.	Female	30

Figure 17 - Event log from an ER process adapted from Park et al., 2023

Case ID	Event ID	Activity	Process type	Timestamp	Originator	Department	Sex	Ages
C10001	…	…	…	…	…	…	…	…
	E1001	Inpatient_drug_1	9202	2021-08-01 09:03:00	O10001	ER dept.	Female	30
	…	…	…	…	…	…	…	…
	E1009	Outpatient_procedure_2	9201	2021-08-01 12:30:00	O10003	ER dept.	Female	30

Figure 18 - Event log from an ER process adapted from Park et al., 2023

On the other hand, patient journey does not appear on any article other than P1 (Park et al., 2023). The nature of this process is different from the rest as it aims to represent the complete medical history of the patients for a certain period. Therefore, the other 3 processes (Inpatient, Outpatient and ER) are fused together to create this process. The authors propose two methods for building the event log, either include every activity from all visits or group them into one activity that represents one single visit as exemplified in figure 18 and 17, respectively. In case the model is based on the event log with detailed information as it is the one from figure 18, it can become excessively complex in terms of the number of branches and overall difficult to analyze, in other words, a "spaghetti" model. The analysis done in P1 (Park et al., 2023) consisted on looking into the most frequent paths in terms of control-flow (ordering of activities) and duration between activities. The author also adds that this kind of process can be used to detect the efficiency of a clinical intervention by comparing the change of the patient's medical condition after the respective treatment.

RQ3:What is the standard pre-processing treatment to the data?

The first stage of making data available for its application in PM is finding the data in its raw format. Several articles (e.g., P15) only refer to Hospital Information System (HIS). However, in article P3 (Pang et al., 2021), there is a multitude of information systems that support a

hospital as it indicates which information systems was data extracted, for instance Radiology Information System (RIS) and Laboratory Information System (LIS). Additionally, other articles refer sources such as Electronic Health Records, from P2 (Dallagassa et al., 2022), or Emergency Department's Electronic Medical Records from P8 (Arias et al., 2020) . However, for processes outside of the medical scope there are other Information Systems with different purposes such as supporting logistics from P4 (Kropp et al., 2023).

The extraction process of the data itself is rarely mentioned. P20 (Ramires & Sampaio, 2022) mentions the extraction of data from a single source of data (SAP) in which the first stage was fetching 3 separate datasheets containing important activities of a procurement process. (Order, Delivery and Consumption). From this point, with the support of SQL, an event log was built where each product (case ID) has 3 rows, respective of each datasheet. However, as it is referred in article P3 (Pang et al., 2021), data can be scattered across several information systems. P1 (Park et al., 2023) refers the use of a Common Data Model (CDM), an informational model that enables the standardization of the format of different databases in terms of relationships and encoding (Hripcsak et al., 2015). More specifically, it refers a CDM adapted to the healthcare context named OMOP-CDM, Observational Medical Outcomes Partnership, which has been applied increasingly across several EHR, according to P1 (Park et al., 2023).

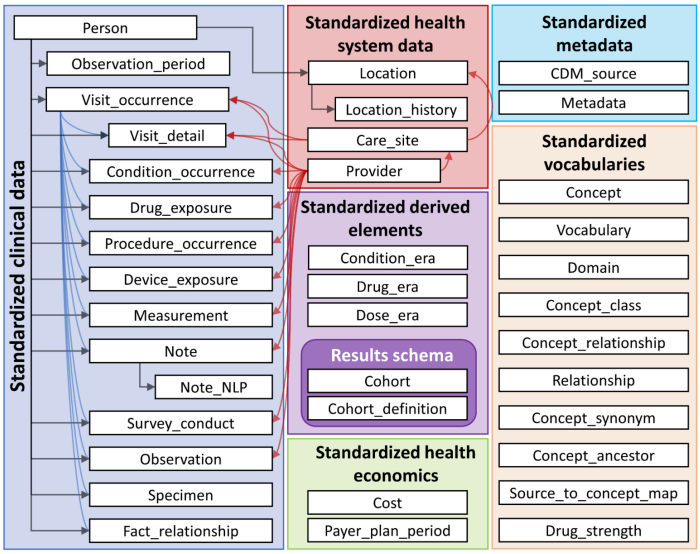

Figure 19 - Overview of the tables in the OMOP-CDM(OHDSI, 2020)

The main purpose of OMOP CDM is to systematize data in a manner that facilitates its analysis. For that reason, the model is considered to be 'person centric', in which all events are linked to a person, as it is shown in figure 19. Additionally, in the same figure it can be seen that there are data elements crucial for an event log such as 'Visit_occurrence' or 'Procedure_occurence' corresponding to the case id and a specific event, respectively (OHDSI, 2020). Article P1 (Park et al., 2023) explains in detail the construction of the event logs through this CDM and the resulting analysis through Process Mining.

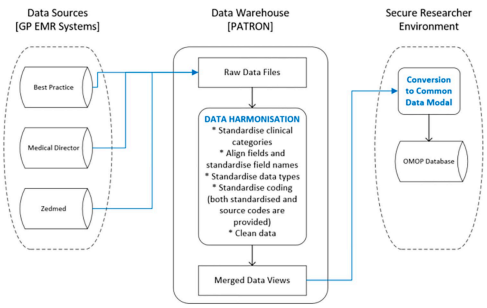

Figure 20 - Harmonization process for EMR sources (Ward et al., 2024)

In figure 20, it is shown a practical instance of a merging process of data from several EMRs (Electronic Medical Records) from 3 vendors (Best Practice, Medical Director, Zedmed) into a single OMOP-based database. The extraction of the raw files from all EMRs was followed by data harmonization, in which data across different sources was standardized through several aspects. More specifically, clinical data such as medical conditions and medication were respectively standardized through internationally recognized terminology SNOMED CT (Systematized Nomenclature of Medicine – Clinical terms) and RxNorm. Additionally, certain aspects of data schema such as data types, field names and the overall format were aligned. Finally, the standardized data was structured according to OMOP CDM and merged into a single database, through an ETL process involving SQL scripts (Ward et al., 2024).

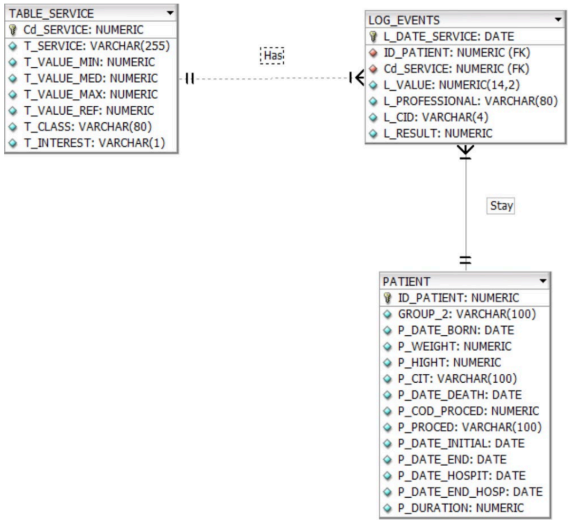

Figure 21 - Data Model of a MBDS adapted from Dallagassa et al., 2022

However, there are instances where the only source of data provides a data model that enables the extraction of the required data for the event log as it is the case in P2 (Dallagassa et al., 2022). In the article, it is mentioned that the Minimum Basic Data Set (MBDS) was provided by the EHR, shown in figure 21. More specifically, table 'LOG_EVENTS' provided information about events (e.g., date, resource) and the associated activities and patients, in tables 'TABLE_SERVICE' and 'PATIENT', respectively.

After the extraction, usually it is also mentioned a layer of data quality which tackles several issues present in the data. Article P18 (Erhard et al., 2023) focus on this topic by enhancing RAMM (Readiness assessment and maturity model), a data maturity model which is a framework focused on discovering the current level of data management. More specifically, it adds a new dimension, 'Event Quality Data' to the current 7 dimensions of RAMM. This dimension has 5 different degrees of maturity which are based on the following criteria for event logs:

- Trustworthy: Fictitious data
- Completeness: Exclusion of real events
- Semantics: Understandability of event's information
- Privacy: Non-disclosure of any sensible data

The most used technique is filtering through several methods that intend mainly on removing inconsistencies (P2, P3, P4, P10, P15) which would otherwise hinder the interpretability of the results. For instance, P2 (Dallagassa et al., 2022) there is a set of rules (e.g., activity transition, event duration) which aim to filter out unusual cases (e.g., Patients' onset time is earlier than arrival). On the other hand, P16 (Munoz-Gama et al., 2022) stresses out the value of infrequent behavior as it can reveal insights into healthcare processes. Additionally, duplicate records were filtered out in both P10 (Agostinelli et al., 2020) and P11 (Rashed et al., 2023). Regarding data granularity, multiple articles (P2, P3, P11, P15) point out that the high level of detail of the data required aggregation of several events of similar type into macro event. Article P3 (Pang et al., 2021) mentions that the excessive variety of events hinders the interpretability of the model, a phenomenon named "Spaghetti effect". With respect to missing data, there is a discrepancy when it comes to dealing with it. In P10 (Agostinelli et al., 2020), it is mentioned that records with missing data are simply erased whereas in P11 (Rashed et al., 2023) it is mentioned that lines are populated instead with the mean values based on the values of other cases. Additionally, in P3(Pang et al., 2021), due to an issue with the lack of timestamps regarding completion dates it was assumed that start and completion timestamps were assumed to be the same for certain activities. Finally, there is also the removal of attributes, in which P4 (Kropp et al., 2023) investigates the usability of attributes through two criteria. The first one is the amount of NULL/NA values present. The other one is the attributes' value variation with time. For instance, attributes such as age or gender do not change over time whereas attributes such as month's information do vary.

Data Privacy is not as studied as other topics with only a few articles (P16, P9) directly addressing it. P16 (Munoz-Gama et al., 2022) introduces the topic of data privacy as one of the key challenges for PM in healthcare due to the use of sensible data from patients.

Table 5 - Level of Fitness for anonymization methods: 'NA': Not Applicable; '+': No impact on PM results,'-': Impact on PM results; '+/-': Impact on certain type of PM results (Pika et al., 2020)

	Attributes				
	Case	**Activity**	**Timestamp**	**Resource**	**Data**
Encryption	+	+/-	+/-	+/-	+/-
Data Swapping	+	-	-	-	-
Noise Addition	-	-	-	-	-
Value Suppression	NA	+/-	+/-	+/-	+/-
Generalization/micro-aggregation	NA	+/-	+/-	+/-	+/-

In article P9 (Pika et al., 2020) , it was laid out several anonymization methods that aim to disclose sensible data, for instance, encryption of attributes inside the event log (e.g., activity, resource). Table 5 gathers the impact in the overall PM results for each data transformation in each attribute. For instance, encryption can be applied to timestamps but it may have a negative impact due to the impossibility of applying performance analysis such as throughput time. Another example is the negative impact of adding noise in all attributes as it will invalidate any PM results. For both value suppression and generalization, it is not logical to apply them in the case attribute for its uniqueness of values. Regarding the rest of the attributes, the method of application can dictate the level of impact. An example of this is the generalization of timestamps which can negatively impact if the order of events is erased, for example, cutting short the date into only its year. Another one is value suppression on infrequent activities which had low impact on the average throughput times but impacted conformance checking on several logs.

Table 6 - Capabilities supported by PM-based software adapted from Rashed et al. (2023) and University of Erlangen-Nürnberg (n.d.)

	ProM	**Disco**	**Celonis**	**PM4PY**
Import support types	MXML, XES	CSV, XLS, MXML, XES, FXL	CSV, XLSX, XES	CSV, XLS, MXML, XES, FXL
Filtering Data	YES	YES	YES	YES

Regarding the format of the event log, it is usually chosen either XES or CSV format which are accepted into the most used PM software in the market (e.g., ProM, Celonis), according to table 6. On a final note, from the same table, it can be concluded that the event log can generally be further transformed through filtering by the software previously referred.

RQ4: What are the most used methodologies?

From the articles that were reviewed only a few directly referred the methodologies used (e.g., paper P7, P10). Additionally, out of the standard methodologies available for Process mining projects (e.g., L∗ life-cycle model or PM²) only one has been referred, more specifically PM².

Nonetheless, other articles have developed their own methodology (P3,P13, P14, P11, P5) for their own goals. For instance, article P14 (Tamburis & Esposito, 2020) develops a methodology which supports the construction of a Discrete Event Simulation (DES) model for any type of process. On the other hand, article P5 (Asare et al., 2020) presents another methodology focused solely on Conformance Checking with event logs from an open-source software, OpenEMR.

Due to the complex nature of healthcare related processes, it is important to have health professionals in the entirety of the project for support as they can give practical insights into the process (Munoz-Gama et al., 2022). The involvement of domain Expertise has been directly mentioned in several articles. In article P3 (Pang et al., 2021), these professionals are called upon to verify the interpretability of the model and work alongside the PM analysts for adjustments in the log filtering or the algorithm itself. Another similar strategy is presented by article P15 (Benevento et al., 2022) which presents the technique IPD (Iterative Process Discovery) for the construction of process models. IPD sets itself aside from the rest of the algorithms associated with process discovery by iteratively placing an activity at a time according to the domain experts and the information from the event logs. The author goes even further and compares the model generated by IPD and models generated by APD (Automated Process Discovery), which includes several algorithms such as heuristic miner without the support of domain experts. The model generated by IPD performed better than the ones generated by APD techniques in several metrics (e.g., precision, fitness, f-score). Likewise, paper P21 (Pereira Detro et al., 2020)also focus on building a process model in combination of event log with domain knowledge through clinical guidelines and inputs from physicians. In addition, article P7 (Pereira et al., 2020) creates a version from the methodology PM² adapted to a healthcare setting, PM²HC, in which one of the pillars from this new methodology is an increased integration between all stakeholders, including health professionals. It argues that in PM² the role of these professionals is not specific, failing to guide readers in integrating them in the overall PM project. As such, across the several stages from this new methodology, which remain the same as the ones from PM², their role is shown in a clear manner. On a final note, even though methodologies are not directly mentioned in the articles, some of concepts may be present. For instance, analysis iteration from PM² seems to be present in P1 (Park et al., 2023) where the clinical pathway was analyzed through different time windows or between the total amount of activities and just a select sample of activities.

RQ5: What are the most used data analysis techniques?

The techniques used amongst the several use cases differs for each focus of analysis. However, there are a set of techniques which set the standard for different purposes.

Table 7 - Distribution of PM-related algorithms through the articles

	Algorithms						
	Inductive	Fuzzy	Alpha	Heuristic	Split	Others	Not specified
Papers	P15,P9,P14,P11,P5,P10	P2,P17,P4	P17,P11	P11	P15	P13,P15	P1,P3,A13,P8,P7

Process discovery, used for generating process models is used by all use cases with the differentiating factor being the algorithm used for each. The most used algorithm was Inductive followed by Fuzzy, showed in table 7. The difference between the different algorithms is rarely addressed with article P17 (Sundari & Nayak, 2022) being the only article that compares the performance between two algorithms (Alpha and Fuzzy) in terms of activity deviation in which the paper found fuzzy miner to perform better.

Variant analysis is used in articles (P1, P4) where several pattern paths in the process model are investigated. In both articles, the focus seems to be on the most common paths where it was compared the ordering of activities. Article P4 (Kropp et al., 2023) goes even further and provides a comparison between variants through a time-based indicator, throughput time.

Another focus of analysis is conformance checking between the event logs and a process model by showing the rate of cases that adhere to the latter (P4, P5). Additionally, in article P4 (Kropp et al., 2023) it is displayed a list of specific violations in the sequence (e.g., activity_x followed by activity_n) with the respective number and percentage of occurrences.

Another important element of analysis is the use of metrics. The most prevalent indicators are time-related such as throughput time, which may be named length of stay or hospitalization days and sojourn time. Article P3 (Pang et al., 2021), uses indicators of a specific pathology (AIS) namely Imaging to needle time (INT) and Door to Needle time (DNT). These time related indicators are especially useful when it comes to perform bottleneck analysis for compliance analysis as in checking if the execution of an activity is done within a certain time frame. Article P2 (Dallagassa et al., 2022) also includes other metrics that can enrich the analysis such as average cost of the treatments and medical-related indicators, more specifically, death rate and relapse rate.

There is also the use of graphs for instance frequency distribution which has been mostly used for bottleneck analysis (e.g., number of patients' dropping out of the hospital per maximum waiting time from article P10 (Agostinelli et al., 2020)). Another example is dotted chart (P1, P10), where each activity of a case can be depicted through a set of dots which eases the analysis in a time perspective. Finally, social network is a visualization used by P10 (Agostinelli et al., 2020) to analyze the role structure of a company as it reveals insights into the relationships between resources (e.g., departments, health professionals).

Paper P19 (Martin et al., 2020) stresses the importance of conducting multi-perspective analysis on the process in order to extract as much information as possible. However, P10 (Agostinelli et al., 2020) is the only article that explicitly addresses the variety of perspectives

of analysis and explores them namely: control-flow, performance and organizational. Control-flow studies the order in which activities are executed. Performance perspective centers around the level of execution of the process, on the efficiency when executing the tasks. Finally, organizational perspective, also known as resource perspective, focuses on resources (e.g., Health professionals, Medical Units) responsible for the execution of activities and the relationship between them (Mannhardt, 2018).

3. METHODOLOGY

3.1. DESIGN SCIENCE RESEARCH

One of the pillars of research in Information System (IS) is Design Science (DS) and its goal is developing the artifacts, which be composed of different elements including software technologies or formal logic and ultimately aim at solving the organizational challenges(Hevner et al., 2004). However, Peffers et al. (2006) argued that at the time of the publication there was a lack of a standardized guiding frameworks from which the DS research could be carried out. Given this, the authors presented a holistic framework named Design Science Research Process (DSRP) which will be the chosen methodology for the upcoming research.

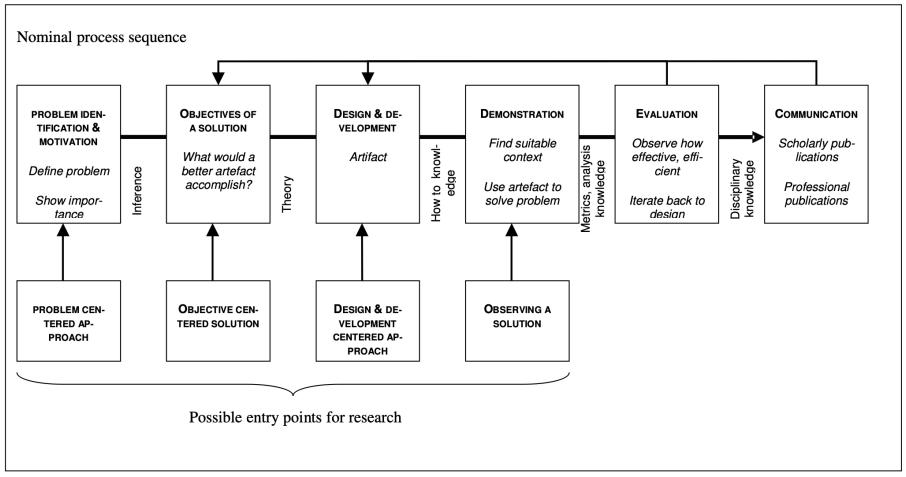

Figure 22 - Overview of DSRP (Peffers et al., 2006)

The development of DSRP model was based on past IS research. More specifically, it was laid out all common research process elements (e.g., Requirements, Development) present in the selected studies which then served as support to create a framework agnostic from any research focus. Given this, the model consists of 6 consecutive tasks but with a certain degree of autonomy as the researcher has the choice to start at any stage depending on the research focus whether that is an objective centered solution or a design and development centered approach, as it is shown in Figure 22.

Figure 22 offers an overview of the entire process in which it is represented through several stages:

- Problem Identification and motivation: Definition of the problem followed by a deep analysis into it to motivate the scientific to pursue solutions.
- Objectives of a solution: Establishment of objectives for the proposed artifact regarding the resolution of the problem referred in the previous stage.

- Design & Development: Development stage of the artifact preceded by its functionality and architecture.
- Demonstration: Application of the developed artifact through a demonstration which can be in the form of a case study or simulation.
- Evaluation: Evaluation of the solution's performance based on the comparison between the results from the demonstration and the objective defined between. In case of insufficient performance, research can retract to the stage of Design and Development to carry out changes to the artifact.
- Communication: The closure of the research by communicating to all interested parties through publishing the information regarding the study, more specifically the research problem and its impact alongside the proposed artifact with its level of performance (Peffers et al., 2006).

3.2. RESEARCH STRATEGY

The structure of this research will be based on several stages of the DSRP model:

- Problem identification and motivation: The problem is explained in sections 1.1 and 1.2. In section 1.1 it is shown a general overview of the challenges which have a direct impact in the health institutions in Portugal. On the other hand, in section 1.2 it is introduces the concept the of Process Mining as tool that can potentially solve the inefficiencies of the hospitals and the ongoing issue with the general lack of knowledge among analysts to apply Process Mining.
- Objectives of a solution: In the face of the insufficient use of data generated by the hospitals, explained in detail in section 1.2, the objectives for this research have been laid out in section 1.3. The key objective for this research is the development of a data architecture that supports the use of Process Mining in a hospital setting.
- Design & Development: Preceding the construction of the architecture (artifact) it was conducted a literature review in section 2 regarding the functioning of an hospital and the main concepts surrounding Process Mining. Additionally, to enrich this chapter, it was also conducted a systematic review through framework PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) to discover the state of the art concerning the use of Process Mining in a healthcare setting. Given the knowledge extracted from the literature review, it is possible to build a data architecture which can support the application of data from a hospital in Process Mining.
- Demonstration: Once the artifact is built, it is shown an instantiation of the architecture containing software alongside the respective details about the functionalities.
- Evaluation: The architecture will be evaluated through a set of interviews with people of preferably different backgrounds in terms of usability in a real-life setting. Additionally, it will be asked for suggestions for improving the architecture and in case

those can be applied in a short timeframe then they can be included in a revised version of the architecture.

- Communication: Presentation of the research to an academic committee and the disposal of the free access of the document for the scientific community through the University's website.

4. EMPIRICAL STUDY

4.1. ASSUMPTIONS

4.1.1. Multidimensional PM Analysis

Based on what was studied in the literature review, about process miming for the health sector, it was possible to build a framework presented in table 8. This table combines several types of processes centered around patients' care and several focus of analysis.

Table 8 - Multidimensions of PM analysis in a hospital setting

	Perspectives b)					
	Control-Flow		Performance		Organizational	
Areas of focus a) / Types of Process c)	Operational	Medical	Operational	Medical	Operational	Medical
ER	X		X		X	
Inpatient	X	X	X	X	X	
Outpatient	X		X		X	
Patient Journey		X		X		
Support	X		X		X	

a) According to research, the focus of analysis can branch off into two viewpoints, operational and medical.

On one hand, there is operation, which is focused on all types of tasks, from administrative to logistics which aim to ultimately ensure the seamless operation within the hospital. It includes all kinds of activities, from clinical (e.g., 'Medication Administration') to non-clinical (e.g., 'Billing'). For this matter, throughout the articles that are several types of analysis which serve this specific purpose:

- Bottleneck analysis through waiting times.
- Optimization of resource (e.g., health professionals) allocation.
- Cost analysis (e.g., average cost of a treatment).

On the other hand, an event log can be built in the perspective of analyzing through a clinical lens in which the purpose is on the on retrieving insights from the patients treatment and their medical condition. Examples of these from past research include:

- Efficiency of the treatment in terms of medical related indicators (e.g., death rate, relapse rate)
- Breakdown of paths (variant analysis) through attributes (e.g., age, gender) or specific activity transition.

On a final note, an event log does not need to be analyzed exclusively through the areas of Operation or Medical. To illustrate this situation, the analysis of a certain process can be composed of questions related to waiting times for each activity (operational) as well as the process variants according to each age group (medical).

b) Data Analysis can be divided into several perspectives:
    - Control Flow concerns the ordering of the activities.
    - Performance examines the efficiency of the execution of the process in terms of certain metrics.
    - Organizational perspective investigates the resources responsible for the execution of the activities.

   Both control-flow and performance can be integrated in both areas of focus, operation and medical. In contrast, organizational perspective is not as relevant to the medical focus as it solely analysis the structure of roles and responsibilities inside the organization.

c) In relation to the processes inside the hospital, it is widely recognized at a high level that there are two types: medical, which is focused on clinical actions, and organizational, which is focused on the support of the clinical actions. However, regarding medical processes, it is also several articles propose a further division, through the type of patients' hospital visit. Therefore, it was established the following types of process:
    - Emergency Room: Visits to the ER

   When only looking into the ER process, the focus tends to be the on operational efficiency due to the time constraint nature of this type of care as certain medical conditions require immediate care, which can be done partly by looking into waiting times of the different activities. As such the level can be of low granularity, for example specific medical actions can be grouped together as a single activity respectively (e.g., medication, examinations) as the clinical details may not be necessary. Nonetheless, ER processes should be more detailed if they are integrated in a wider inpatient process as the focus of analysis can also be medical and specific information regarding the different medical actions is relevant.

    - Outpatient: Visit with a duration of less than a day

   Outpatient processes are generally characterized by lasting up until one day and no hospitalization of the patients. From past research, this process is usually composed of clinical and non-clinical activities and only explored in an operational perspective where examples of analysis include, waiting times for consultations or examinations and social networks regarding the dynamics between resources.

    - Inpatient: Visit that lasts more than a day

From previous research, inpatient process can be analyzed through both an operational and medical viewpoint. Regarding medical focus, with inpatient processes, different treatments can be compared in terms of medical indicators such as death rate, relapse rate and even control-flow and length of stay (LOS) for different groups of patients based on the available attributes, for instance age or gender. On the other hand, one can also focus on the bottlenecks or waiting times that patients may face during their treatment or the costs associated with the activities.

- Patient Journey: Historical data which includes all sorts of hospital visits

Patient journey differs from the rest as its temporal focus varies greatly as it can analyze data from years ago up until this moment. Given the existence of relatively old data, it may be not feasible to analyze it in an operational standpoint (Daniel et al., 2008), for instance data from 3 or 4 years ago is not valid when analyzing current bottlenecks. However, the timeframe of a scientific research can extend for several years (Bebu et al., 2017) as such the value of data for the medical focus maintains for longer. The mentioned analysis for this kind of process was variant analysis through the breakdown of the most frequent paths and the effectiveness of a treatment through the patient's medical condition.

- Support: Support to the functioning of the hospital

Finally, support process, which is mostly known as organizational process, due to its nature, is analyzed solely on an operational scope.

4.1.2. Methodology

Relating to the methodology, the key takeaway from past research is based on two premises:

- Actively involving domain experts in several stages. It is required for both technical analysts and domain experts to be aligned in what concerns initial and later stages of the project. In the beginning, input of domain experts is crucial when designing the research questions. Likewise, in the end it is crucial their presence in order to interpret the results.
- Iteration between data processing and analysis stage until the research questions defined in the beginning stage are answered. From the research, there is the possibility of repeatedly filtering the event log to fetch different process models with a different outlook on the process.

4.1.3. Extraction and Transformation of Data

The sources from which one can extract data are usually information systems adapted to a medical setting, commonly known as Hospital System Information (HIS) including:

- Eletronic Medical Record (EMR)
- Electronic Health Record (EHR)

- Radiology Information System (RIS)
- Laboratory Information System (LIS)

Besides HIS, data can be extracted through other sources such as information systems focused on logistics.

The foundation for the construction of event logs is the rule that all events should be linked to a specific case. The choice of the case rests in the type of process to be investigated. For instance, if it is considered the division of healthcare processes described previously (e.g.,) then the chosen case should correspond to each type of visit, in other words, columns such as 'visit_id'. However, events can be scattered around different information systems. In case this situation occurs then there are available CDMs specific to healthcare such as OMOP-CDM which support the merging process regarding:

- Database structure: Single data structure in terms of entities (tables), primary/foreign keys, relationship between entities.
- Terminology: Conversion of clinical terms such as medical condition or medication in respect to a certain standard clinically related terminology (e.g., SNOMED CT)
- Data Format: Align the structure of the data itself (e.g., conversion of all date columns into the format 'YYYYMMDD')
- Data Types: Align the data types of all columns (e.g., date columns should be of string type)

Once these issues are taken care of then they can be merged into one single database with a data model that enables the creation of the event logs.

Once the data is finally merged and with all the relevant events and the case linked, then it is the stage of creating the event log. There is no general method to create the event log from a database, however the starting point should be to flatten the database into one view by ensuring that data containing the events is tied to a case. Afterwards, apply the necessary transformations through an ETL tool so that the final table resembles one of an event log. The output can then be exported to a wide variety of file formats such as XES, CSV or XLS for the subsequent import by the PM-based software. After the extraction of the event log from the sources there is still a certain amount of work required when it comes to data quality and data privacy.

4.1.4. Data Quality and Data Privacy

Data quality of the event logs, according to the use cases analyzed previously, does not only revolve around the level of accuracy when representing reality but also how easy it is to analyze it. Given this, in the face of certain issues, one can proceed with specific actions:

- Inconsistent/Outlier behavior: The decision to include inconsistent or infrequent behavior comes down to the decision of the stakeholders in the project. There is an

argument to be had to remove this kind of data if the objective is to uncover the main behavioral pattern, for instance through the inclusion filters based on arbitrary rules (e.g., activity #1 should be followed by activity #2). However, by filtering out the noise there might be a loss of informational value as it could reveal insights into its root causes.

- Granularity change: High granular data results in highly complex process models that might come as a challenge to analyze. An instance of this may be the phenomenon of spaghetti models that are hard to interpret and originated by event logs with high-detailed activities. Therefore, one has the option to aggregate events into macro-events. An example of this is aggregating both events 'Administration of drug X' or 'Administration of drug Y' into 'Drug administration'.
- Missing data: Events can have different missing elements such as timestamps, activity names or missing cases, which result in different data treatments. The fastest method is to filter out cases or events missing essential data such as timestamps or case ids. Another method can be to populate with the support of statistical values (e.g., mean).
- Removal of attributes: Raw data may have attributes which have to be deleted. One reason for it can be the fact that it just does not have enough number of proper entries or it is not relevant for the focus of analysis.
- Duplicated data: Duplicated records should be deleted.

Regarding data privacy, during the extraction of data from the source, there may be precautionary actions on not disclosing sensible data. If the latter situation does not happen, then there is an array of data transformation techniques which can be applied into each event log's attribute in order to preserve data privacy. Ultimately, the objective from this stage is to prevent the identification of patients through their records.

Table 9 - Application of data transformation techniques for each events log's attribute

	Attributes				
	Case	Activity	Timestamp	Resource	Data
Encryption	X	X	X	X	X
Value Suppression		X	X	X	X
Generalization/Micro-Aggregation		X	X	X	X

Table 9 shows which techniques can be applied for which attributes. For both Value Suppression and Generalization/Micro-Aggregation, it is not feasible to apply them in the case attribute as it is necessary for each case to have a unique value. Encryption, on the other hand, can be applied in all instances.

However, each method has its own drawbacks as there is always a degree of loss of informational value when applying them. For instance, suppressing infrequent activities (activity suppression) can have impact in the results, for example in conformance checking where event log is being compared with a process model.

4.1.5. Process Mining techniques

Finally for data analysis, table 10 places the main techniques into the different perspectives according to their objective.

Table 10 - Allocation of the PM techniques to the different perspectives

Techniques \ Perspectives		Control-Flow	Performance	Organizational
Process Discovery		X		X
Conformance Checking		X		
Variant Analysis		X		
Metrics			X	
Visualizations	Frequency Distribution	X	X	X
	Social Network			X

The different perspectives have a range of instruments of support:

- Process Discovery: Generation of process models represented by visual notations such as Petri-Net or DFG (Directly-Follows Graph) which focus on control-flow with the capture of event sequencing. However, it may help with organizational perspective as certain algorithms can produce BPMN diagrams which may show the resource behind execution of each event.
- Conformance checking: Comparison between event logs and an arbitrary process model in terms of event sequencing.
- Variant Analysis: Holistic analysis into the different pattern paths or sequence of events.
- Metrics: Indicators that can assess the process through several lenses. It includes time-based indicators such as throughput time, sojourn time as well as other indicators (e.g., cost, medical conditions).
- Visualizations: Information regarding the process can be displayed through graphical plots:
    - Frequency distributions can apply to all perspectives including:
        - Dotted charts displaying points representing an activity of each case over a time span enabling the analysis of activity ordering (control-flow perspective) and bottleneck analysis (performance perspective)
        - Distribution of activities assigned to the different resources (organizational perspective)
    - Social network: Resources are represented by nodes and connection lines representing the relationship between them, as such it is considered solely part of the organizational perspective.

4.2. ARCHITECTURE FOR THE APPLICATION OF PROCESS MINING IN AN HOSPITAL SETTING

The proposed architecture is depicted below in figure 23 and it cover several stages that were addressed in the previous chapter. From scope identification to the different type of Process Mining techniques, this architecture gives an overview of the requirements to create a process mining project in a hospital.

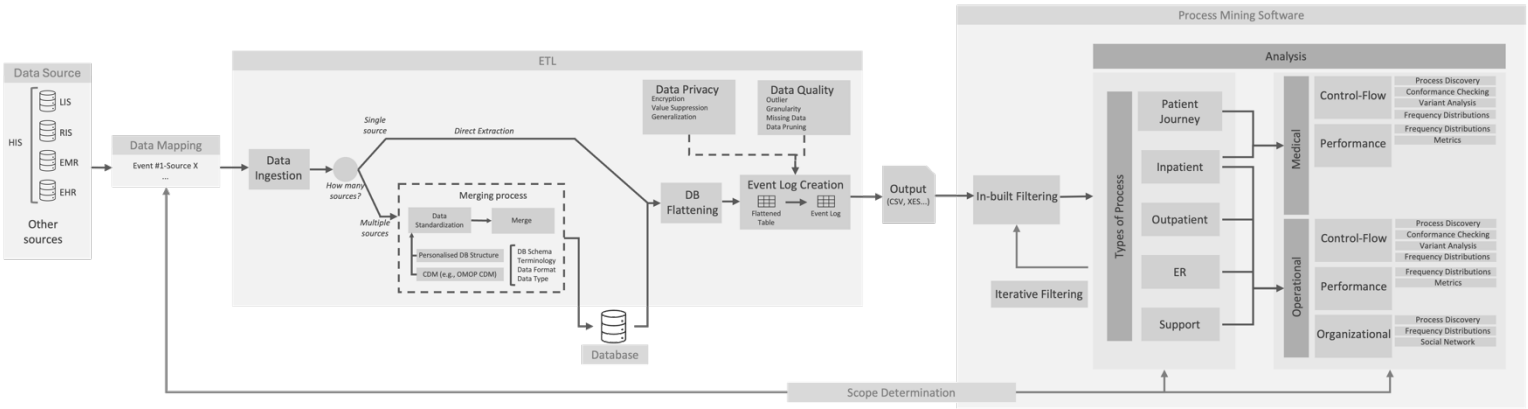

Figure 23 - Proposal for the data architecture

Below the architecture is described in detail.

4.2.1. Scope Identification

This is the first stage of any Process Mining project and aims at defining essentially what is being investigated and the type of conclusions to take from it. For this matter, both domain experts (e.g., administrators, health professionals) and technical experts (e.g., PM analysts) must cooperate to align the following key questions:

- What process to investigate?

The discussion must consider the different type of processes in a hospital setting. First, there are those processes centered around the path that patients take around the hospital in which the division is made through the type of visit (ER, Outpatient, Inpatient and Patient Journey). Then, there are support processes which act as secondary processes for the functioning of a hospital (e.g., billing, patient transportation). Important to refer that the choice of the process is totally dependent on the availability of the case id. For that reason a layer of data mapping should be required as it allows to track the location of the relevant events in the respective sources. Additionally, certain activities that make up the process might be scattered across several information systems and other type of sources, which may pose a challenge and further increase the workload.

- What should the analysis address?

At this stage the research questions (e.g., 'What is the waiting time for the urgency care triage?') should be formulated according to the type of process that were chosen in

conjunction to the area of focus whether it should lean towards to the functioning of the hospital (Operational area) or more focused on clinical research (medical area).

4.2.2. ETL

Covers the transformation process from raw data in the source to the final Event Log. This phase might change depending on the number of sources from where data needs to be extracted from:

- For single source, it is only required to flatten the relationships of the database with a key column in consideration, namely, the case id.
- In case of multiple sources, the scattered data requires to be merged into a single database. For support purposes, one can decide to choose a Common Data Model (CDM) such as OMOP CDM. The required data transformations are the following:
    - Data Schema: Definition of the tables and the respective relationships.
    - Terminology: Conversion of clinical terms into a standardized code (e.g., SNOMED CT).
    - Data Type: Align data type (e.g., string)
    - Data Format: Align data format (e.g., date formats)

  Once the merging database is built, one can proceed to create a flatten view as described previously.

The flattened table allows the view of all possible events available in the database without the need to conduct joins. Given this, the flattened table has all the data required in the event log and one can finally proceed with its creation. During this stage, data transformations at the format level of the flattened table should be conducted to fulfill all necessary specifications of an event log. This stage covers several steps which can happen simultaneously depending on the project's context, therefore there is a certain amount of autonomy on how to plan it.

- Data Privacy: In case of need, there are available several data privacy methods that can be applied in all attributes, for instance:
    - Encryption
    - Value Suppression
    - Generalization/Micro-Aggregation

  Important to stress that Encryption is the only acceptable method to anonymize case attribute as the other two would erase the uniqueness nature of each case id.

- Data Quality: Facilitate the analysis of the process model generated by the software through several tools:
    - Aggregation of events
    - Removal of duplicated records.
    - Removal of outliers.

    - Populate missing data with statistical values (e.g., mean)
    - Removal of attributes not relevant for analysis or with a reduced number of proper entries

After all transformations are done and the event log is built, the output should be converted into any format that is acceptable for the PM-based software, such as CSV or XES, depending on the ETL capabilities.

4.2.3. Process Mining Software

Once the data is ingested, there may exist additional filtering capabilities when it comes to manipulate event logs before application of PM techniques (e.g., Process Discovery).

Finally, depending on the research questions, which were determined in the first stage (Scope Determination), there is a range of capabilities offered by PM-based software:

- Control-Flow: Analysis of the behavior of the process through the sequence of the activities.
    - Process Discovery
    - Conformance Checking
    - Variant Analysis
    - Frequency Distribution
- Performance: Evaluation of performance of the process in terms of several criteria including time (e.g., bottleneck analysis), cost or medical results (e.g., relapse rate, mortality rate):
    - Metrics
    - Frequency Distribution
- Organizational: Analysis of the resources (e.g., health professionals, departments) which executed the activities:
    - Process Discovery
    - Social Network Mining
    - Frequency Distribution

In case, the research questions, defined during scope determination, have not been answered and the software has filtering capabilities then there may be the option to iteratively modify the event log until all of the research questions are answered (Iterative filtering)

4.3. DEMONSTRATION

In this chapter, it is presented an instantiation of the proposed architecture where it is shown several software products that are available in the market, as it is show below in figure 24. It is shown how the software products fit in the different modules and stages of the architecture, from the source to the data analysis.

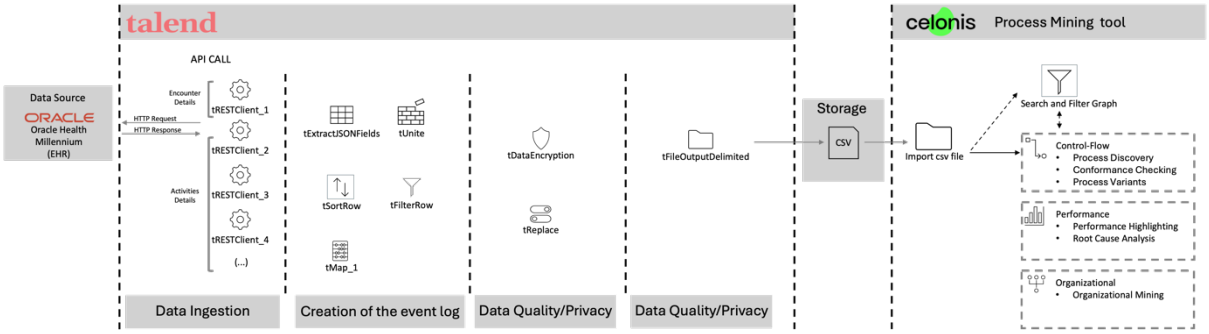

Figure 24 - Instantiation of the proposed data architecture

The source, Oracle Cerner Millennium, is an Electronic Health Record (EHS), a type of HIS with centralized information regarding the patient, from observations to lab records.

Data transformation phase is executed through an ETL tool (Extract, transform and load) named Talend Studio. This software has a wide array of capabilities that can handle all the required capabilities for data transformation, from its raw state until its final form, ready to be exported into the process mining software. Regarding the Process Mining software, it was chosen for this use case one of the most used named Celonis.

The extraction phase can be made through API requests that fetch json files containing data that enables the creation an event log which can recreate a patient's path in hospital. This is a snippet into the metadata of the files available in Oracle Cerner Millennium:

- Encounter: Overview of the patient's visit, which include fields such as:
  - *ID:* Contains Encounter ID
  - *Class*: Type of visit (e.g., inpatient, outpatient)
  - *Subject*: Patient's ID and Name
  - *Period*: Start/End time of the visit
  - …
- Observation: Checking of vital signs, laboratory results, diagnostic reports
  - *Encounter:* Contains Encounter ID
  - *Category:* Identification of the observation
  - *EffectiveDateTime*: Execution date
  - …
- Procedure: Details regarding clinical procedures
  - *Encounter:* Contains Encounter ID
  - *Code:* Identification of the observation (e.g., Vital Signs)
  - *Subject: Identification of the patient*
  - *Status: Execution status (e.g., completed)*
  - *PerformedDateTime:* Execution date
  - …

Besides these, there are other examples such as appointments, medication requests and even nutrition orders. But most importantly, these files possess a key (Encounter Id) that represents the specific visit to the hospital (Oracle Health Millennium, n.d.). Therefore, it can be guaranteed that events are all related to a case id.

The following paragraph describes in theory some of the components available for the construction of the event log in Talend Studio, the ETL tool. As such it will only display a high-level explanation of the components and will exclude any other specificities of building a pipeline. This software has capabilities regarding the ingestion of data from HIS and data transformation techniques required for formatting tables into the event log (Talend, n.d.) including:

- tRESTClient: Direct import of JSON files into Talend Studio, through API calls.
- tExtractJSONFields: Execution of transformations in the JSON file such as flattening all nested fields and rename them.
- tFilterRow: Filtering of rows according to certain conditions.
- tMAP: Application of several type of joins.
- tUnite: Union of tables.
- tSortRow: Sort of rows through ids and dates.

Additionally, it has components which cover questions of data privacy and data quality:

- tDataEncrypt: Application of encryption methods into any column.
- tReplace: Replacement of certain values, which can be used as a tool of aggregation of activities.

In the final step, the software has the capability to export the final table, in the event log format, into a wide variety of formats:

- Tfileoutputdelimited: Export of the final table into a csv file to a pre-determined path, whether that is a local or cloud storage (Talend, n.d.).

Celonis, comparatively to other PM-based software options such as ProM, is a much more user friendly tool, which represents a gentler learning curve. In addition, it has a wide variety of capabilities that focus on every perspective referred earlier (Control-Flow, Performance and Control-Flow). Another capability worth mentioning is the cross over between the different perspectives. For instance, it is possible to analyze cases that do not conform to a certain arbitrary process model (Conformance Checking) in terms of throughput time (time-based indicator). On a final note, there are also filtering capabilities through several angles (e.g., transition of activities) which can manipulate the rest of the analysis automatically(University of Erlangen-Nürnberg, n.d.).

4.4. EVALUATION AND DISCUSSION

In order to validate the utility of the architecture it was made two interviews with professionals of the Information systems from the area of health more specifically from the Portuguese healthcare system. Table 11 gives an overview into the background of each participant. The interviews did not have a fixed structure but it followed to certain extent the following steps:

1. Presentation of the architecture and its several stages.
2. Discussion:
    - Pros and Cons of the architecture
    - Recommendations

Table 11 - Background of the interviewed participants

Interviewer ID	Background	Domain
E1	Specialist in Systems and Information Technologies	Industry
E2	Director of the service of Systems and Information Technologies and Communication/Head of IT department	Industry

The main point referred by both interviewees was the vast range of sources of data within a hospital. Participant E1 refers that even though most data regarding the patients is centralized in a database (SONHO), which is transversal across the Portuguese public healthcare system, it still leaves out a certain amount of data from other information systems (e.g., SClínico). Given this, both participants agree that the proposed architecture should represent the different sources (e.g., RIS, LIS). E2 goes even further on this topic in the sense that with the variety of sources, it is required a layer of standardization of the data before merging it as it is seen in the proposed architecture. However, in a more practical sense, E1 referred that the architecture can improve by showing all available sources and the respective type of data (e.g., exams, prescriptions).

Along the interview, E2 focused on the transformations made within the event log, more specifically data quality and privacy. Regarding data privacy, there are already tools for it when data is requested for the purpose of statistical analysis. However, it is given a special attention in anonymizing data related to rarer diagnosis in order to prevent the identification of the patients. Additionally, E2 raised several challenges for data quality including the constant incomplete filling of data in information systems as well as the removal of inconsistent data as unusual behavior can be the result of a clinical decision which may be important for analysis.

Finally, regarding the stage of Process Mining, E1 had very limited remarks regarding the range of processes and the respective type of PM analysis. E1 argues that the capabilities of Process Mining are directed towards the administration of the Hospital than to IT staff. Additionally,

E2 argues that the people in charge of analyzing the medical area are mostly composed professionals other than physicians (e.g., nurses, pharmacists), therefore, for a broader term, the area of focus should be named clinical instead of medical. Nonetheless, E2 recognizes the overall value of the last part of the architecture through two lenses. The first is the fact that the framework can depict the different types of processes as well as showing the respective eligible analysis. The second is the inclusion of metrics which can help visualize bottlenecks in the hospital services. E2 stresses the fact that there is an increasing pressure in these services which then results in significant waiting lists.

5. CONCLUSIONS

5.1. SYNTHESIS OF THE DEVELOPED WORK

The thesis was developed based on the set of stages from DSRP. First it was established the problem and motivation which were based on the increasing pressure on the Portuguese health services which demanded a solution which contributed to the efficiency of the processes inside the hospital. The solution, which was decided in the objective stage, was to develop an architecture that could give an overview on how to apply process mining in any hospital. In order to build the architecture, it was necessary to have strong foundations of the main concepts of process mining, functioning of the hospital and a current situation of process Mining in the healthcare setting through a systematic review. Given the knowledge obtained in the literature review it was possible to build an architecture with all important steps, from the various data sources to the available Process Mining techniques. Afterwards, for demonstration purposes, it was done an instantiation of the architecture using real-life software for the various stages (e.g., Celonis, Talend). Finally, the architecture was evaluated through two interviews with specialists who provided valuable insights into the architecture in terms of pros and cons and recommendations.

5.2. LIMITATIONS

The first limitation of this work was the lack of interviews due to the lack of time. The interviews did not have much focus on the process mining stage of the architecture due to the background of the experts. Another limitation due to the lack of time, is the lack of testing in a real-life context in a hospital. The last limitation is somewhat related with the previous one as the high-level nature of the architecture does not allow for more detailed information regarding the specificities of the information systems or type of processes within each hospital.

5.3. FUTURE WORK

The future set of actions should be on conducting several tests on the architecture in order to improve it even further. More interviews should be conducted with a special focus in the Process Mining stage. These interviews should include professionals that partake in both administration of the hospital as well as in the clinical research. The second method should be the application of the architecture within a real hospital. These two tests should reveal potential improvements for the effectiveness of the architecture.

BIBLIOGRAPHICAL REFERENCES

AbbVie, & Nova IMS. (2023). *ÍNDICE DE SAÚDE SUSTENTÁVEL 2022/23*. https://www.abbvie.pt/content/dam/abbvie-dotcom/pt/Documents/Índice%20Saúde%20Sustentável%202023.pdf

Agostinelli, S., Covino, F., D'Agnese, G., De Crea, C., Leotta, F., & Marrella, A. (2020). Supporting Governance in Healthcare Through Process Mining: A Case Study. *IEEE Access*, *8*, 186012–186025. https://doi.org/10.1109/ACCESS.2020.3030318

Arias, M., Rojas, E., Aguirre, S., Cornejo, F., Munoz-Gama, J., Sepúlveda, M., & Capurro, D. (2020). Mapping the Patient's Journey in Healthcare through Process Mining. *International Journal of Environmental Research and Public Health*, *17*(18), 6586. https://doi.org/10.3390/ijerph17186586

Asare, E., Wang, L., & Fang, X. (2020). Conformance Checking: Workflow of Hospitals and Workflow of Open-Source EMRs. *IEEE Access*, *8*, 139546–139566. https://doi.org/10.1109/ACCESS.2020.3012147

Bebu, I., Braffett, B. H., Pop-Busui, R., Orchard, T. J., Nathan, D. M., & Lachin, J. M. (2017). The relationship of blood glucose with cardiovascular disease is mediated over time by traditional risk factors in type 1 diabetes: the DCCT/EDIC study. *Diabetologia*, *60*(10), 2084–2091. https://doi.org/10.1007/s00125-017-4374-4

Ben Sassi, S., & Yanes, N. (2023). *Data Science in Healthcare: a Systematic Review*. https://doi.org/10.13140/RG.2.2.32836.19842

Benevento, E., Aloini, D., & van der Aalst, W. M. P. (2022). How Can Interactive Process Discovery Address Data Quality Issues in Real Business Settings? Evidence from a Case Study in Healthcare. *Journal of Biomedical Informatics*, *130*, 104083. https://doi.org/10.1016/j.jbi.2022.104083

Bovonratwet, P., Webb, M. L., Ondeck, N. T., Lukasiewicz, A. M., Cui, J. J., McLynn, R. P., & Grauer, J. N. (2017). Definitional Differences of "Outpatient" Versus "Inpatient" THA and TKA Can Affect Study Outcomes. *Clinical Orthopaedics and Related Research*, *475*(12), 2917–2925. https://doi.org/10.1007/s11999-017-5236-6

Caldwell, P. H., & Bennett, T. (2020). Easy guide to conducting a systematic review History of Systematic Reviews Why Do a Systematic Review? *Journal of Paediatrics and Child Health*, *56*, 853–856. https://doi.org/10.1111/jpc.14853

Cancer Diagnostic Delay Reduction. Retrieved March 18, 2024, from https://www.tf-pm.org/resources/casestudy/cancer-diagnostic-delay-reduction

Centers for Disease Control and Prevention. (2023). *Outpatient visit*. https://www.cdc.gov/nchs/hus/sources-definitions/outpatient-visit.htm

Centro Hospitalar Lisboa Norte, E. (2016). *Relatório e contas* . https://www.ulssm.min-saude.pt/media/k2/attachments/administracao/Relatorio%20e%20Contas%202016.pdf

Conselho das Finanças Públicas. (2024). *SECTOR EMPRESARIAL DO ESTADO 2021-2022*.

Dallagassa, M. R., Iachecen, F., Furlan, L. H. P., Ioshii, S. O., & Carvalho, D. R. (2022). Applying process mining in health technology assessment. *Health and Technology*, *12*(5), 931–941. https://doi.org/10.1007/s12553-022-00692-5

Daniel, F., Casati, F., Palpanas, T., & Chayka, O. (2008). *Managing Data Quality in Business Intelligence Applications*.

De Roock, E., & Martin, N. (2022). Process mining in healthcare – An updated perspective on the state of the art. *Journal of Biomedical Informatics*, *127*, 103995. https://doi.org/10.1016/J.JBI.2022.103995

Erhard, A., Arthofer, K., & Helm, E. (2023). *Extending a Data Management Maturity Model for Process Mining in Healthcare*. https://doi.org/10.3233/SHTI230038

Helm, E., Lin, A. M., Baumgartner, D., Lin, A. C., & Küng, J. (2020). Towards the Use of Standardized Terms in Clinical Case Studies for Process Mining in Healthcare. *International Journal of Environmental Research and Public Health*, *17*(4). https://doi.org/10.3390/ijerph17041348

Hevner, March, Park, & Ram. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75. https://doi.org/10.2307/25148625

Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., Suchard, M. A., Park, R. W., Wong, I. C. K., Rijnbeek, P. R., van der Lei, J., Pratt, N., Norén, G. N., Li, Y.-C., Stang, P. E., Madigan, D., & Ryan, P. B. (2015). Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Studies in Health Technology and Informatics*, *216*, 574–578.

Instituto Nacional de Estatística. (2022). *Censos 2021- XVI Recenseamento Geral da População. VI Recenseamento Geral da Habitação*. https://www.ine.pt/xurl/pub/65586079

Kropp, T., Faeghi, S., & Lennerts, K. (2023). Evaluation of patient transport service in hospitals using process mining methods: Patients' perspective. *The International Journal of Health Planning and Management*, *38*(2), 430–456. https://doi.org/10.1002/hpm.3593

Kruse, C. S., Goswamy, R., Raval, Y., & Marawi, S. (2016). Challenges and Opportunities of Big Data in Health Care: A Systematic Review. *JMIR Medical Informatics*, *4*(4), e38. https://doi.org/10.2196/medinform.5359

Lenz, R., & Reichert, M. (2007). IT support for healthcare processes – premises, challenges, perspectives. *Data & Knowledge Engineering*, *61*(1), 39–58. https://doi.org/10.1016/J.DATAK.2006.04.007

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of Clinical Epidemiology*, *62*(10), e1–e34. https://doi.org/10.1016/J.JCLINEPI.2009.06.006

Lillrank, P., & Liukko, M. (2004). Standard, routine and non-routine processes in health care. *International Journal of Health Care Quality Assurance*, *17*(1), 39–46. https://doi.org/10.1108/09526860410515927

Mannhardt, F. (2018). *Multi-perspective Process Mining*.

Manole, F., Marian, P., Mekeres, G. M., & Voiţă-Mekereş, F. (2023). Systematic Review of the Effect of Aging on Health Costs. *Archives of Pharmacy Practice*, *14*(3), 58–61. https://doi.org/10.51847/npqdV19MYv

Mans, R. S., Aalst, W. van der, & Vanwersch, R. J. B. (2015). *Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes*. Springer Publishing Company, Incorporated.

Martin, N., De Weerdt, J., Fernández-Llatas, C., Gal, A., Gatta, R., Ibáñez, G., Johnson, O., Mannhardt, F., Marco-Ruiz, L., Mertens, S., Munoz-Gama, J., Seoane, F., Vanthienen, J., Wynn, M. T., Boilève, D. B., Bergs, J., Joosten-Melis, M., Schretlen, S., & Van Acker, B. (2020). Recommendations for enhancing the usability and understandability of process mining in healthcare. *Artificial Intelligence in Medicine*, *109*, 101962. https://doi.org/10.1016/j.artmed.2020.101962

Massachusetts Health Policy Commission. (2014). *2013 Cost Trends Report*.

Munoz-Gama, J., Martin, N., Fernandez-Llatas, C., Johnson, O. A., Sepúlveda, M., Helm, E., Galvez-Yanjari, V., Rojas, E., Martinez-Millana, A., Aloini, D., Amantea, I. A., Andrews, R., Arias, M., Beerepoot, I., Benevento, E., Burattin, A., Capurro, D., Carmona, J., Comuzzi, M., … Zerbato, F. (2022a). Process mining for healthcare: Characteristics and challenges. *Journal of Biomedical Informatics*, *127*, 103994. https://doi.org/10.1016/J.JBI.2022.103994

Munoz-Gama, J., Martin, N., Fernandez-Llatas, C., Johnson, O. A., Sepúlveda, M., Helm, E., Galvez-Yanjari, V., Rojas, E., Martinez-Millana, A., Aloini, D., Amantea, I. A., Andrews, R., Arias, M., Beerepoot, I., Benevento, E., Burattin, A., Capurro, D., Carmona, J., Comuzzi, M., … Zerbato, F. (2022b). Process mining for healthcare: Characteristics and challenges.

Journal of Biomedical Informatics, *127*, 103994. https://doi.org/10.1016/j.jbi.2022.103994

OHDSI. (2020). The Book of OHDSI. In *https://ohdsi.github.io/TheBookOfOhdsi/*. https://ohdsi.github.io/TheBookOfOhdsi/

Oracle Health Millennium. (n.d.). *FHIR R4 APIs for Oracle Health Millennium Platform*. Https://Docs.Oracle.Com/En/Industries/Health/Millennium-Platform-Apis/Mfrap/Op-Observation-Id-Get.Html.

Ostroff, F. (1999). *The Horizontal Organization: What the Organization of the Future Actually Looks Like and How It Delivers Value to Customers*. Oxford University Press. https://books.google.pt/books?id=BWXnBwAAQBAJ

Pang, J., Xu, H., Ren, J., Yang, J., Li, M., Lu, D., & Zhao, D. (2021). Process mining framework with time perspective for understanding acute care: a case study of AIS in hospitals. *BMC Medical Informatics and Decision Making*, *21*(1), 354. https://doi.org/10.1186/s12911-021-01725-1

Park, K., Cho, M., Song, M., Yoo, S., Baek, H., Kim, S., & Kim, K. (2023). Exploring the potential of OMOP common data model for process mining in healthcare. *PloS One*, *18*(1), e0279641. https://doi.org/10.1371/journal.pone.0279641

Peffers, K., Tuunanen, T., Gengler, C., & Rossi, M. (2006). The design science research process: a model for producing and presenting information systems research. *Proceedings Design Research Information Systems and Technology DESRIST'06*, *24*.

Pereira Detro, S., Santos, E. A. P., Panetto, H., Loures, E. De, Lezoche, M., & Cabral Moro Barra, C. (2020). Applying process mining and semantic reasoning for process model customisation in healthcare. *Enterprise Information Systems*, *14*(7), 983–1009. https://doi.org/10.1080/17517575.2019.1632382

Pereira, G. B., Santos, E. A. P., & Maceno, M. M. C. (2020). Process mining project methodology in healthcare: a case study in a tertiary hospital. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *9*(1), 28. https://doi.org/10.1007/s13721-020-00227-w

Pika, A., Wynn, M. T., Budiono, S., ter Hofstede, A. H. M., van der Aalst, W. M. P., & Reijers, H. A. (2020). Privacy-Preserving Process Mining in Healthcare. *International Journal of Environmental Research and Public Health*, *17*(5), 1612. https://doi.org/10.3390/ijerph17051612

Ramires, F., & Sampaio, P. (2022). Process mining and lean six sigma: a novel approach to analyze the supply chain quality of a hospital. *International Journal of Lean Six Sigma*, *13*(3), 594–621. https://doi.org/10.1108/IJLSS-12-2020-0226

Rashed, A.-H. M., El-Attar, N. E., Salama Abdelminaam, D., & Abdelfatah, M. (2023). PROCESS MINING APPROACH FOR DISCOVERING AND ANALYZING THE HEALTHCARE PROCESSES IN PYTHON. *Journal of Theoretical and Applied Information Technology*, *31*(16). www.jatit.org

Rojas, E., Munoz-Gama, J., Sepúlveda, M., & Capurro, D. (2016a). Process mining in healthcare: A literature review. *Journal of Biomedical Informatics*, *61*, 224–236. https://doi.org/10.1016/j.jbi.2016.04.007

Rojas, E., Munoz-Gama, J., Sepúlveda, M., & Capurro, D. (2016b). Process mining in healthcare: A literature review. *Journal of Biomedical Informatics*, *61*, 224–236. https://doi.org/10.1016/j.jbi.2016.04.007

Sousa, P. A. F. de. (2009). O sistema de saúde em Portugal: realizações e desafios. *Acta Paulista de Enfermagem*, *22*.

Subrahmanya, S. V. G., Shetty, D. K., Patil, V., Hameed, B. M. Z., Paul, R., Smriti, K., Naik, N., & Somani, B. K. (2022). The role of data science in healthcare advancements: applications, benefits, and future prospects. *Irish Journal of Medical Science (1971 -)*, *191*(4), 1473–1483. https://doi.org/10.1007/s11845-021-02730-z

Sundari, M. S., & Nayak, R. (2022). *Performance Analysis and Activity Deviation Discovery in Event Log Using Process Mining Tool for Hospital System*.

Talend. (n.d.). *About Talend components*. Https://Help.Talend.Com/En-US/Components/8.0/Content/Components/Home.Htm.

Tamburis, O., & Esposito, C. (2020). Process mining as support to simulation modeling: A hospital-based case study. *Simulation Modelling Practice and Theory*, *104*, 102149. https://doi.org/10.1016/j.simpat.2020.102149

Theis, J., Galanter, W. L., Boyd, A. D., & Darabi, H. (2022). Improving the In-Hospital Mortality Prediction of Diabetes ICU Patients Using a Process Mining/Deep Learning Architecture. *IEEE Journal of Biomedical and Health Informatics*, *26*(1), 388–399. https://doi.org/10.1109/JBHI.2021.3092969

Toussaint, J. S., & Berry, L. L. (2013). The Promise of Lean in Health Care. *Mayo Clinic Proceedings*, *88*(1), 74–82. https://doi.org/10.1016/j.mayocp.2012.07.025

University of Erlangen-Nürnberg. (n.d.). *Celonis Process Mining*. Https://Www.Processmining-Software.Com/Tools/Celonis-Process-Mining.

Valero-Ramon, Z., Fernandez-Llatas, C., Valdivieso, B., & Traver, V. (2020). Dynamic Models Supporting Personalised Chronic Disease Management through Healthcare Sensors with Interactive Process Mining. *Sensors*, *20*(18), 5330. https://doi.org/10.3390/s20185330

van der Aalst, W. (2016). *Process Mining*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-49851-4

van der Aalst, W., Adriansyah, A., de Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., Bose, J. C., van den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., de Leoni, M., … Wynn, M. (2012). *Process Mining Manifesto* (pp. 169–194). https://doi.org/10.1007/978-3-642-28108-2_19

van der Aalst, W., Adriansyah, A., & van Dongen, B. (2012). Replaying history on process models for conformance checking and performance analysis. *WIREs Data Mining and Knowledge Discovery*, *2*(2), 182–192. https://doi.org/10.1002/widm.1045

van der Aalst, W. M. P. (2011). *Process Mining*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-19345-3

van Eck, M. L., Lu, X., Leemans, S. J. J., & van der Aalst, W. M. P. (2015). *PM^2: A Process Mining Project Methodology* (pp. 297–313). https://doi.org/10.1007/978-3-319-19069-3_19

Ward, R., Hallinan, C. M., Ormiston-Smith, D., Chidgey, C., & Boyle, D. (2024). The OMOP common data model in Australian primary care data: Building a quality research ready harmonised dataset. *PLOS ONE*, *19*(4), e0301557. https://doi.org/10.1371/journal.pone.0301557

Zimmermann, L., Zerbato, F., & Weber, B. (2023). What makes life for process mining analysts difficult? A reflection of challenges. *Software and Systems Modeling*. https://doi.org/10.1007/s10270-023-01134-0

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa