



DAVID DUQUE DELGADINHO

BSc in Computer Science and Engineering

ANALYSIS OF ELECTRONIC HEALTH RECORDS TO ASSESS COMPLIANCE WITH CLINICAL PATHWAYS

MASTER IN COMPUTER SCIENCE AND ENGINEERING

NOVA University Lisbon
September, 2024



ANALYSIS OF ELECTRONIC HEALTH RECORDS TO ASSESS COMPLIANCE WITH CLINICAL PATHWAYS

DAVID DUQUE DELGADINHO

BSc in Computer Science and Engineering

Adviser: Daniel Catita Martins
Software Engineer, UpHill Health

Co-adviser: Joaquim Ferreira da Silva
Assistant Professor, NOVA University Lisbon

Analysis of electronic health records to assess compliance with Clinical Pathways

Copyright © David Duque Delgadinho, NOVA School of Science and Technology, NOVA University Lisbon.

The NOVA School of Science and Technology and the NOVA University Lisbon have the right, perpetual and without geographical boundaries, to file and publish this dissertation through printed copies reproduced on paper or on digital form, or by any other means known or that may be invented, and to disseminate through scientific repositories and admit its copying and distribution for non-commercial, educational or research purposes, as long as credit is given to the author and editor.

ACKNOWLEDGEMENTS

First of all, I would like to express my gratitude to all the people who have helped me throughout these five years at the University and who have allowed me to be here completing this journey.

Thank you to my supervisor, Professor Joaquim Silva, for all his help throughout the development of the thesis, all the calls, all the feedback and all the information that enabled me to deliver this thesis today.

I would also like to thank Daniel Catita and Rodrigo Pio for their tireless support throughout the months, for the endless calls, the timely clarifications, the friendly advice and, above all, for answering my questions for the 4th, 5th and 6th time, always with a smile on their faces. A special thanks to everyone who works at UpHill, especially the Engineering and Medical teams for lifting me up on the bad days and keeping me afloat on the good ones and more importantly, for teaching me that with hard work and perseverance, I can achieve anything that I set my mind to.

To my friends, thank you for always being there to lift my spirits, for your patience when I disappeared for weeks, and for your understanding during stressful times. A special thank you to Ricardo Romão, Jaime Saramago, Alexandre Oliveira, Bernardo Silva, Tiago Henriques and Romão Costa for all the memorable moments we shared, the projects, the study sessions and the late-night discussions, for helping me at all times and for making this journey much more enjoyable. To Afonso Gama, for the continuous support every single day, for giving me the motivation to not give up on the bad days, to all the lessons learned along the way, all the past stories and the ones that are yet to come. And to Miguel Paiva, for the rides, the parties, the jokes, the mockings, the arguments, but most importantly, the friendship.

Last, but not least, to my family. To my parents for the love, for always believing in me, always pushing me to be better, always giving me the freedom to pursue my own goals and for being there every step of the way. To my brother, for giving me the guidance and the knowledge, for your faith and constant support, for teaching me the little things, and most importantly, for being the example of the type of Man I want to become one day.

Thank you!

ABSTRACT

An efficient and high quality healthcare delivery is based on well evaluated clinical approaches that meet the individual needs of each patient and the effective utilization of resources by the healthcare organization. Clinical Pathways (CPWs) are the standardized plans in healthcare focused on management of specific medical issues or procedures that allow establishment of consistent high quality of service delivery. Considering that Clinical Processes (CPs) are very complex, it is very much possible that deviations arise between that process carried by the healthcare organization and the standardized CPW, which ultimately affect the quality of the service provided and, potentially, the clinical outcomes. Therefore, it is vital to evaluate this compliance in order to determine the quality of the clinical processes being developed and allow an optimization of it to, consequently, improve the efficiency in healthcare delivery.

This thesis proposes a strategy to analyze CPs stored in a Clinical Data Repository (CDR) and measure their compliance with standardized CPWs. The main objectives were identifying clinically relevant indicators which demonstrated the efficiency of clinical journeys and measuring the organizational resource usage. This was done by carefully outlining non-compliance situations and implementing metrics that allowed me to draw the necessary conclusions. For this to be possible, DocumentDB was integrated with real-time streaming to enable the capture of changes in the CDR to ensure that compliance analysis was carried out with the most up-to-date information and Extract, Transform, Load (ETL) processes were utilized to transform and load this data into the database, making sure that the strategy used was as efficient and scalable as possible.

Keywords: Clinical Processes, Clinical Pathways, Compliance Assessment, Real-time Streaming, DocumentDB, ETL

RESUMO

Uma prestação de cuidados de saúde eficiente e de elevada qualidade baseia-se em abordagens clínicas bem avaliadas que satisfazem as necessidades individuais de cada doente e na utilização eficaz dos recursos por parte da organização de cuidados de saúde. Os Pathways Clínicos (PWCs) são os planos padronizados nos cuidados de saúde centrados na gestão de problemas ou procedimentos médicos específicos que permitem estabelecer uma prestação de serviços consistente e de elevada qualidade. Considerando que os Processos Clínicos (PCs) são muito complexos, é muito possível que surjam desvios entre esse processo realizado pela organização de cuidados de saúde e o Clinical Pathway padronizado, o que acaba por afetar a qualidade do serviço prestado e, potencialmente, os resultados clínicos. Assim, é fundamental avaliar esta conformidade para determinar a qualidade dos processos clínicos que estão a ser desenvolvidos e permitir a sua otimização para, consequentemente, melhorar a eficiência na prestação de cuidados de saúde.

Esta tese propõe uma estratégia para analisar os PCs armazenados num Repositório de Dados Clínicos (RDC) e medir a sua conformidade com os PCWs padronizados. Os principais objectivos foram identificar indicadores clinicamente relevantes que demonstrassem a eficiência dos percursos clínicos e medir a utilização dos recursos organizacionais. Para tal, foi necessário delinear cuidadosamente as situações de incumprimento e implementar métricas que me permitissem tirar as conclusões necessárias. Para que tal fosse possível, DocumentDB foi integrado com streaming em tempo real para permitir a captura de alterações no CDR, de modo a garantir que a análise de conformidade fosse efectuada com a informação mais actualizada e foram utilizados processos de Extract, Transform, Load (ETL) para transformar e carregar estes dados na base de dados, assegurando que a estratégia utilizada fosse tão eficiente e escalável quanto possível.

Palavras-chave: Processos Clínicos, Pathways Clínicos, Avaliação da conformidade, Streaming em tempo real, DocumentDB, ETL

CONTENTS

List of Figures	vii
List of Tables	viii
Acronyms	ix
1 Introduction	1
1.1 Motivation	1
1.2 Goals	2
1.3 Company Framework	3
1.4 Document Structure	4
2 State of the Art	5
2.1 Clinical processes	5
2.1.1 Clinical Pathways	6
2.1.2 Compliance Checking and Monitoring	9
2.1.3 Evaluating Compliance to Clinical Pathways	10
2.2 Data Analytics Processes	11
2.2.1 Real-Time Streaming	11
2.2.2 Extract, Transform, Load	13
2.2.3 Power BI	14
2.3 Data Analytics Databases	15
2.3.1 Relational vs Non-Relational Databases	16
2.3.2 MongoDB	17
2.3.3 DocumentDB	19
2.3.4 MySQL	20
2.4 Clinical Data Repository	22
2.4.1 Definition	22
2.4.2 Healthcare Data Standards	23
2.4.3 From HL7v2 to CDR	26

3	Solution	28
3.1	Methodology	28
3.2	Solution	29
3.2.1	Clinical Data Repository	29
3.2.2	Change Data Capture	29
3.2.3	Change-stream-service	31
3.2.4	Deidentification-service	31
3.2.5	Pending-resource-repository	32
3.2.6	Compliance-service	33
3.2.7	Analytics	34
3.2.8	Dashboard	39
4	Results	41
4.1	Data processing	41
4.2	Metrics	41
4.2.1	General Compliance	41
4.2.2	Compliance by Action of the pathway	44
4.2.3	Compliance by Phase of the pathway	46
4.2.4	Compliance per Week	48
5	Conclusion and Future Work	50
5.1	Conclusion	50
5.2	Future Work	51
	Bibliography	52

LIST OF FIGURES

2.1	Example of an Oncology Clinical Pathway (CPW)	8
2.2	Document example	20
2.3	mongoDB vs MySQL	21
2.4	HL7v2 message example	24
2.5	FHIR message example	26
2.6	FHIR Mapping	26
3.1	Solution	30
3.2	PowerBI workspace	40
4.1	Fitness values by compliance levels of entire clinical pathway.	44
4.2	Fitness values by compliance levels of steps with interoperability.	45
4.3	Compliance values by Action of the pathway	46
4.4	Compliance values by Phase of the entire clinical pathway.	47
4.5	Compliance values by Phase of the steps with interoperability.	47
4.6	Compliance values per Week of the entire clinical pathway.	49
4.7	Compliance values per Week of the steps with interoperability.	49

LIST OF TABLES

2.1	Relational vs Non-Relational	18
4.1	Overall Data for entire clinical pathway.	42
4.2	Overall Data of steps with interoperability.	42

ACRONYMS

CDR	Clinical Data Repository (<i>pp.</i> 3 , 5 , 22 , 26 , 29)
CP	Clinical Process (<i>pp.</i> 1 , 4–6 , 9 , 22 , 23 , 27 , 33)
CPW	Clinical Pathway (<i>pp.</i> vii , 1–5 , 7–11 , 22 , 27 , 31 , 33 , 35 , 38)
EHR	Electronic Health Records (<i>pp.</i> 5 , 14 , 17 , 19 , 22)
FHIR	Fast Healthcare Interoperability Resources (<i>pp.</i> 25 , 29)

INTRODUCTION

1.1 Motivation

The process of choosing clinical approaches carefully and applying them when needed is instrumental in defining the healthcare delivery scenario. This decision-making procedure not only affects the organizational integrity but has a deep impact on transparency and effectiveness of clinical practices, which in turn influence patients' experience within the healthcare system. In this sense, healthcare processes must be highly dynamic, inclusive and multidisciplinary, enabling the automation and sequential programming of standard activities for the treatment of multiple pathologies. Among the vast possibilities of strategies that can be adopted, clinical protocols, commonly known as Clinical Pathways (CPWs) are an effective way for improving standardized and high-quality care.

CPWs are noteworthy for promoting the quality of healthcare and standardizing patient care by providing an organized, evidence-based framework. They are structured plans that outline evidence-based guidelines and interventions to standardize patient care for specific medical conditions or procedures. These protocols are carefully developed to assist healthcare personnel in finding the best way of treating and managing certain clinical diseases. As presented by [60], a CPW have the following features:

1. adapts clinical recommendations into institution-specific frameworks;
2. outlines the sequence of actions required in treatment following a structured protocol;
3. seeks to standardize care for a particular clinical issue.

In many domains, when a process is carried out, differences are noted between the anticipated and actual performance, which may also affect the outcomes. In healthcare services this happens a lot and the Clinical Process (CP) of healthcare institutions often deviate from the standardized CPWs [62]. This deviation is extremely important to contain and techniques need to be assembled in order to shrink this deviation, to make sure the best treatment is always being delivered. In recent years, hospital information systems

have developed rapidly, becoming more complex and more capable of managing the growing amount of data generated. Simultaneously, with the advancements in science and technology, the demand for higher quality delivery and customized service also increases [72]. From this point of view, it is becoming increasingly important to analyse the clinical practices observed and evaluate their compliance with the pre-established CPW in order to improve the protocols implemented, reduce variations that are considered unnecessary and adopt new techniques to enforce best practice, with the ultimate goal of providing a high-quality health service [26].

By establishing CPWs as the structuring of good practices in the definition and execution of care plans for the treatment of clinical cases, it is possible to establish comparative parameters that confirm the compliance of data compiled from users' electronic health records (EHRs) in relation to the processes described in the respective CPWs. As said by [77], managing compliance is crucial for organizations striving to achieve greater process maturity. Additionally, [77] said that once non-compliant cases of operations are identified, the healthcare institutions need to update their processes or use new mechanisms to improve their execution.

Adding all this with the possibility of reaching this compliance checking in real-time, we are able to add a layer of responsiveness and adaptability to the healthcare system, by facilitating rapid modifications and enhancements to care plans in response to changing patient demands and insights obtained from patient data.

1.2 Goals

The primary objective of this thesis is to present a strategy to analyse EHRs via real-time streaming, which are stored in a CDR-FIM (Clinical Data Repository in FHIR Information Model) and measure their compliance with CPWs, with the help of previously established metrics. This way, healthcare organisations can effectively analyse and check for deviations in a patient journey thanks to the information gathered and, consequently, optimize their clinical workflows in order to achieve better patient outcomes.

The following were the main goals established for this thesis:

1. Establishment of clinically relevant indicators that demonstrate the effectiveness of the results of clinical journeys when crossed with their respective compliance or non-compliance with CPWs;
2. Establishment of relevant indicators from an organizational point of view to assess the effectiveness and rationalization of the use of human and physical resources when cross-referenced with their respective compliance or non-compliance with CPWs;
3. Categorization of the different levels of non-compliance that can occur in a CP.

4. Technical and functional analysis of a real-time streaming solution for gathering electronic health records within the outlined scope, retrieved from a Clinical Data Repository (CDR) standardized on the basis of the FIM (FHIR Information Model, a central component of the HL7 FHIR);
5. Development of a solution for retrieving and analysing data derived from the streaming process to support the real-time calculation of the indicators previously established in real time, applying the statistical and analytical methods identified;
6. Automated and continuous evaluation of the results obtained by the processes described above, as well as the quality of the data obtained based on the previously established criteria.
7. Construction of indicators using interactive data visualization tools through Power BI.

In this thesis, the compliance checking will be done in a specific field of Oncology, in the colorectal cancer.

1.3 Company Framework

This thesis is being conducted in collaboration with UpHill, a company focused on digital services for health, with products that integrate with existing hospital information systems to automate tasks and fill gaps in healthcare. UpHill develops CPWs and pathway-based clinical decision support systems (CDSS). They work with leading healthcare providers and life sciences companies throughout Europe to reduce waste related to care variability and fragmentation, while increasing patient outcomes and revenue.

With UPHILL orchestrating care, hospitals have a comprehensive view of their patients' condition throughout their care journeys as individuals and as a population, including past history and expected next steps.

UPHILL helps healthcare teams increase capacity and reduce waiting lists by:

1. Automating tasks such as scheduling exams and appointments (after approval);
2. Monitoring patients' signs and symptoms throughout their entire clinical journey, generating alerts;
3. Patient stratification, so that doctors focus on the patients who need them;
4. Obtaining relevant records automatically from the Electronic Health Record System;
5. Sharing educational materials and best practice recommendations with patients.

The software is applicable to various hospital areas, from the emergency room to the operating room. To accelerate value delivery during deployment, UpHill offers a clinical

journey template library. Based on the most up-to-date research, these templates provide a framework that can be customised to the particularities of various units and are verified by their clinical teams.

These CPWs will be used in this thesis as standard to evaluate the compliance of CPs. This thesis research contributes to UpHill's ongoing efforts to help healthcare teams with their clinical decisions, optimize resource allocation, and deliver value-driven solutions that positively impact patient care delivery across various hospital areas.

1.4 Document Structure

The remainder of this document is organised as follows:

- **Chapter 2:** This is the State of the art. This is where the most relevant content for this thesis is presented to enable a better understanding of the system. It is divided into a more theoretical part where the concepts more related to medicine are clarified and a more technical part where the different components used throughout the solution are explained.
- **Chapter 3:** This is the Proposed Solution and is divided into 2 topics. The first explains the methodology behind the solution. The second outlines, step by step, the solution achieved with this thesis and the different components needed to achieve it;
- **Chapter 4:** This is the result section for the colorectal cancer compliance assessment. The first topic presents the 2 studies used to evaluate our results. The second topic show the results of each metric that was previously established. It starts with the overall results of the compliance mechanism and then dives into the more specific metrics that were done;
- **Chapter 5:** Summarizes the key contributions of this work and talks about the future work that can be developed in this area.

STATE OF THE ART

This chapter introduces the essential concepts for this thesis, where every important component is explained so that the reader who is less familiar with the area of study can understand what is being presented.

The first topic covers the more theoretical side of the thesis, where key medical concepts like EHRs, CPs, CPWs, and compliance between them are clarified. While this section delves into a more medical area rather than a technological one, it is crucial to understand these concepts to better grasp the goal of the thesis and the usefulness of the tools discussed in later topics. The next three topics delve into the more technical aspects of the thesis. The second topic explores data analytics processes in healthcare technology, including real-time streaming, extract, transform, load (ETL) techniques, and Power BI. The third topic defines the importance of storing information safely and provides not only an overview of the advantages of different types of databases (relational and non-relational) but also a more specific view of the databases to be used in this thesis (DocumentDB and MySQL). Finally, the fourth topic defines CDRs and explores the role of healthcare data standards, such as HL7v2 and FHIR, and the importance of interoperability in their development and implementation.

2.1 Clinical processes

Electronic Health Records (EHR)s are key to healthcare analytics, providing a simple and efficient way for healthcare providers and organizations to share data [36], using the healthcare data standards mentioned in the previous topic.

While EHRs are electronic systems used to store and manage patients' health information, clinical processes are the systematic workflows followed by healthcare providers to deliver care to patients, as soon as someone is admitted to the healthcare facility and has his first evaluation and diagnosis, to the treatment of the disease identified, the follow-up, and continued supervision. The components of a CP usually involve patient assessment and diagnosis, treatment plan development and implementation with constant monitoring, and evaluation, as well as comprehensive documentation [39], all

this ensuring a commitment to quality improvement.

Some of the CPs characteristics are the following [7]:

- **Organized Approach:** follow a defined sequence of steps, ensuring that care is delivered in a structured and predictable manner.
- **Patient is the priority:** prioritize the necessities of the patient, securing that care is customised to their individual circumstances and preferences.
- **Collaboration between peers:** coordinated effort from various healthcare professionals, including physicians, nurses, pharmacists, and other specialists.
- **Continuous Improvement:** subject to ongoing evaluation and improvement efforts to adapt to advancements in medicine, technology, and patient needs.

By ensuring consistency and accountability, CPs goal is to enhance the overall capability and performance in the delivery of healthcare, leading to improved patient outcomes.

Optimizing CPs results in a number of advantages, including reduced medical errors, more efficient workflows, and more effective use of resources. According to [35], most of the errors done in healthcare facilities are caused by flawed systems, rather than individual mistakes. So, improving healthcare processes may have a significant effect on patients' quality of life. However, there are always a number of obstacles to overcome, making improvements challenging. Key challenges include the requirement to cut service expenses, minimize patient waiting periods, and increase the productivity of resources while ensuring greater transparency in processes [59]. Although the CPs are executed accordingly with effective universally approved clinical guidelines [62], the assessment of those challenges needs to be addressed, separately, by each organization.

CPs can have different applications, such as oncology and cardiology, where its established the diagnosis, treatment and management of those diseases; surgery, where the preoperative, interoperate, and postoperative care of surgical patients is standardized; and primary care, where healthcare professionals guide the comprehensive care of patients with various health concerns.

The clinical approach to oncology disease management is a comprehensive healthcare specialty that requires complex decision-making [57] that typically involves several stages, including screening and diagnosis, staging, treatment planning, treatment implementation, and follow-up care. The specific steps may vary based on the type of cancer, individual patient characteristics, and advancements in medical knowledge and technology.

2.1.1 Clinical Pathways

The complex field of healthcare needs to be evolving at all times. With all the current problems in hospitals and medical institutions, with long queues and a lack of staff, more

than ever, the process of delivering healthcare has to be quick and efficient. Despite this, in order to achieve quickness and efficiency, we can not give up the correctness and certainty that the best care is being given to everyone. This requires a systematic and coordinated approach that aligns medical interventions with the best available evidence and resources. In response to this need, CPWs have emerged as indispensable tools.

A Clinical Pathway, also known as a care pathway or clinical care pathway, is a multidisciplinary management tool that outlines the order and timing of healthcare interventions related to a particular diagnosis, procedure, or condition. It is designed to standardize the care process to ensure that patients receive appropriate and timely care, based on best practices and evidence-based guidelines [48]. These pathways provide a structured approach to patient care, incorporating key clinical activities, interventions, and expected outcomes at different points along the care journey. [24] described CPWs as an essential structural component aimed at minimizing deviations in healthcare while enhancing the predictability, measurability, and accountability of costs and outcomes.

Figure 2.1 shows an example of a pathway where it can be seen the different steps that need to be followed when treating a disease, allong with the different possible scenarios of treatment. It is obvious that the implementation of these CPWs carry an increased effort, but this is compensated with the increased efficiency in healthcare delivery by offering a clear roadmap for decision-making, optimizing workflows, and minimizing delays in interventions. CPWs are viewed as a way of standardising and making processes more transparent, reducing variability and, consequently, reducing the probability of medical errors.

The protocol also contains the sequence and duration of each action included in the treatment, optimising the results of the therapy implemented and maximising the efficiency of the institution [61]. This way, there is also an improvement in communication with patients, promoting access to the information contained in the care plan to be implemented in a clear manner. Additionally, the ability of the CPW to make it possible to identify risks and anticipate intervention promotes patient safety, increasing its satisfaction with the care provided. However, it should be emphasised that the process described in any CPW is not mandatory, as it is not always the most suitable for all patients, and the clinical manager has the autonomy to apply the treatments they deem necessary and most appropriate for each patient's individual situation, even if these actions differ from what is advised in the CPW. However, the indications in a CPW exist because, according to solid scientific studies, they increase the quality of the service provided, the efficiency of treatments and the utilization of necessary resources.

Standardization and implementation of evidence-based practices are key aspects to the healthcare industry, which technology is often an important component in helping to accomplish. Despite this, treatment regimens within pathways must be regularly evaluated and adjusted to ensure they include the most current and high-quality options [22].



Figure 2.1: Example of an Oncology CPW

2.1.2 Compliance Checking and Monitoring

Given the increasing global deployment of CPWs, compliance checking is becoming more and more critical to the healthcare setting [47]. Compliance checking and monitoring of CPWs are crucial for ensuring the successful execution of standardized CPWs. These systematic processes involve continuous evaluation to confirm that healthcare providers adhere to the specified protocols, interventions, and timelines outlined in the CPW [38].

Practically, what happens is that hospitals collect and store patient's demographics and clinical information, including the treatment behaviours conducted by the medical team and all his treatment journey. Then there are the CPWs, which, as explained in the last topic, are the standardized care procedures which the practitioners are expected to follow. By comparing each event of the CP to the corresponding event in the CPW, it becomes feasible to evaluate adherence to the anticipated trajectory of care.

Most of the compliance mechanisms used focus on aggregated data viewed from an external perspective, such as length of stay, cost and infection rate. These metrics often provide only a retrospective view of overall conformance with established CPs, making it difficult to thoroughly and almost instantly check for compliance with the essential treatment behaviors of CPs [38].

Adhering to internationally recognised principles and best practices is essential for reducing costs [65], transitioning to high-quality, value-based healthcare [3], improve patient outcomes [80] and allowing for more consistency in care treatments throughout the world [38], ensuring an increase in healthcare efficiency and quality.

There are various ways that we can evaluate compliance of CPs to CPWs. Some of them are:

- **Performance Metrics** - Using quantitative measures and key performance indicators (KPIs) to assess the success and compliance of healthcare providers with the CPW. Example: Tracking metrics such as adherence rates, patient satisfaction scores, and complication rates as indicators of pathway compliance.
- **Process Adherence** - Assessing whether healthcare providers follow the specified processes and procedures outlined in CPWs or care protocols. Example: Verifying that the sequence of diagnostic tests and treatment interventions is adhered to according to the CPW.
- **Timeliness of Interventions** - Evaluating whether interventions and treatments are performed within the designated timeframes as specified in CPWs. Example: Ensuring that critical medications are administered promptly according to the recommended schedule.
- **Outcome Tracking** - Monitoring patient outcomes to assess whether the CPW or care plan is achieving its intended goals. Example: analysing data on treatment

response, complications, and other relevant outcomes to measure the success of the implemented interventions.

These types of compliance checking mechanisms contribute to the overall success of CPWs, ensuring that healthcare practices align with evidence-based guidelines and best practices. The selection of specific types of compliance checking depends on the goals of the CPW, the nature of the healthcare setting, and the characteristics of the patient population.

Additionally, as it will be shown in Section 2.2.1, achieving real-time compliance between clinical processes and clinical pathways implies aligning operational procedures and treatment protocols with established standards and regulations in healthcare delivery.

2.1.3 Evaluating Compliance to Clinical Pathways

Clinical pathways analysis has been extensively researched, particularly in the area of healthcare management, where the potential for information extraction has led to a growing interest in gaining insights into actual clinical practice.

In [38], they use a semantic rule-based CPW compliance system designed to monitor critical treatment behaviors in real-time. The system is built on a Clinical Pathway Ontology (CPO) model, which formalizes compliance checking by encoding treatment constraints as Semantic Web Rule Language (SWRL) rules. These rules capture the temporal and behavioral structure of a CPW and are integrated into the clinical process. This allows for ongoing compliance evaluation during the execution of the CPW, providing clinicians with timely alerts and recommendations to address any deviations from the pathway. A case study on the angina pathway in a cardiology department demonstrates the system's effectiveness in detecting non-compliant behaviors and supporting clinicians in making informed decisions to improve pathway adherence in real-time.

The paper [1] discusses a method for evaluating the compliance between observed behavior and a model of expected behavior in a process. It introduces the concept of fitness, which measures the degree of compliance, and proposes an algorithm to construct the best matching instance. The approach involves analysing deviations, such as skipped or inserted activities, and uses cost functions to quantify fitness. The document presents experiments to demonstrate the effectiveness of the approach in handling deviations and provides comparisons with classical fitness measurements. It also discusses the influence of cost ratios on fitness values in the presence of deviations. This particular form of compliance will be used in our solution, as will be explained later.

In [62], the researchers explore the use of process mining techniques to compare clinical pathways and actual clinical practice. They applied declarative models, which offer flexibility suited for the unpredictable nature of healthcare processes. In a case study at the urology department of the Isala hospital in the Netherlands, they used process mining to analyse event data from clinical executions and compare it with initial models based on clinical pathways. By improving these models with real-world data, they

identified deviations and assessed conformance to pathways. Using tools like ProM and Declare, the study demonstrated that process mining is a good approach for enriching models with diagnostic insights and ensuring alignment between clinical practice and best-practice pathways.

Also in the field of process mining, a very common method to analyse the compliance is the Business Process Modeling Notation (BPMN). In [5], they use BPMNs to check the conformance between open-source Electronic Medical Records (EMR) systems and hospital workflows. The study uses BPMNs to model hospital processes and converts them into Petri nets. Event logs extracted from the EMR system are processed and analysed using the ProM Framework for conformance checking. By comparing actual workflow executions in OpenEMR with hospital-defined workflows, the authors assess compliance based on metrics such as fitness, precision, simplicity, and generalization. The results indicate that the processes within the OpenEMR system align with hospital workflows, confirming the utility of open-source EMRs in maintaining compliance with healthcare institutions' operational standards.

Finally, highlighting how much these compliance mechanisms are important in the field of healthcare, there is also work being done in terms of legal rules being applied in the spread of patient information. In [2], the study presents a semi-automated methodology for regulatory compliance checking in healthcare, focusing on the Hospital at Home (HaH) service. The approach integrates Business Process Management and Notation (BPMN) with Defeasible Deontic Logic (DDL) to formalize rules, particularly those related to the European General Data Protection Regulation (GDPR). By modeling hospital processes and encoding GDPR provisions into a compliance checker called Regorous, the system can automatically verify whether business processes, such as patient data handling, meet regulatory requirements. The tool identifies non-compliance by tracing obligations and permissions in processes, offering a real-time solution to monitor and maintain legal compliance in healthcare services. This framework enhances the ability to align healthcare operations with evolving regulations, improving patient care while ensuring adherence to legal standards.

2.2 Data Analytics Processes

2.2.1 Real-Time Streaming

Analysing the compliance of EHRs with their respective CPWs is extremely important for any hospital, as it allows them to understand what, if anything, is being done wrong in a patient's clinical journey, and to be able to adapt that journey according to the results obtained. For this process to be optimized, it is crucial that compliance is done with the most up-to-date clinical information, in order to obtain quality data and make the process as reliable as possible [21].

Real-time streaming involves continuous data transmission from one source to another

in real time so that information is obtained almost instantaneously as it arrives. This technology is applied in several sectors, including the medical field where timely access to current data can enable early detection of the condition [14].

According to [66], there are several rules that must be followed by every stream processing system, including:

1. **Continuous Data Flow:** stream processing systems should handle data in real-time without relying on costly storage operations that introduce unnecessary latency;
2. **Manage Stream Irregularities:** have fault-tolerance mechanisms against stream "irregularities";
3. **Integrate Stored and Streaming Data:** use a uniform language, such as embedded databases, in order to integrate state information with live streaming data;
4. **Process and Respond Instantaneously:** handle high data volumes with minimal latency by optimizing execution paths and integrating key functions into a single process, ensuring high performance with minimal overhead.

Real-time insights enable healthcare professionals to detect anomalies and timely respond to changes in patient status, ultimately leading to enhanced care delivery quality and operational efficiency [67]. This real-time access to clinical data is also a game-changer for compliance validation efforts, providing organizations with opportunities to, not only see, but fix those possible problems. This guarantees that the compliance checks, validation processes and analytical insights rely on the latest and most precise data.

The normal method of data management usually includes batch processing, in which data is accumulated over time and then loaded into a database periodically (every day at 9am, for example). Although this method has worked successfully, especially in cases involving large volumes of data [67], it presents several limitations, especially in scenarios where data accuracy are critical, as in the case of CDRs (Section 2.4), where the output can be substantially delayed [67]. So, when transitioning from the standard batch processing to real-time streaming, there are some steps that require deep focus in order to maintain data correctness, such as the implementation of data validation mechanisms to guarantee data integrity and accuracy, and the implementation of security and privacy measures to protect patient data, such as anonymization.

Despite all this, it is always essential to ensure a good tradeoff between cost and performance [68]. While real-time streaming may offer advantages in terms of scalability and performance, it may come with higher initial costs and complexity. Even with this, when it comes to checking compliance, streaming the CDR in real time turns out to be the ideal choice, as it, once again, ensures that the most updated data is always being analysed.

2.2.2 Extract, Transform, Load

Having data from different heterogeneous sources can be very difficult in terms of processing it. Also, as the volume of data increases, it becomes impossible for conventional methods to check the data quality and accuracy. So, it becomes imperative that new mechanisms are implemented to solve that. Extract, Transform, and Load (ETL) process offers a structure to standardize and combine different data types into a unified local repository [73].

Extract, transform and load (ETL) involves pulling data from multiple sources, adjusting it to meet operational requirements and storing it in a designated structure [73]. ETL plays a vital role in organizing data and enabling this data to be analysed according to each organizations needs.

To make things more clear, a deep dive into each component is essential to understand the advantage of this process.

2.2.2.1 Extract

The Extract is the first phase of this process and it involves linking diverse data sources and retrieving the information needed. Handling a variety of data sources demands specialized tools, due the approaches that each of them can have for connecting and extracting information [34]. Before extraction, is also important to implement a data selection criteria in order to extract only the relevant data from each source, filtering out unnecessary information. Information can be extracted in many ways, such as, by full extraction (extracting all data), incremental extraction (extracting only new or modified data), or change data capture (tracking changes in source data in real-time).

Change-data-capture is a particular interesting approach due to its ability to combine the process of extracting information with the implementation of real-time streaming. The CDC system monitors the database for changes, and then sends these changes to a processing system responsible for the second phase of the ETL, the Transformation phase.

2.2.2.2 Transform

Here, the information that was retrieved in various sources is checked, cleaned and conformed [52]. The transformation phase, as the name suggests, requires converting the extracted data into a standardized format for the subsequent phase [34]. The idea is to ensure consistency, quality, and compatibility with the target system. Transformation tasks may include [37]:

- **Data Quality Checking:** Detecting and addressing quality issues like missing entries or duplicate records;
- **Data Aggregation:** Summarize and aggregate the data to gather helpful insights, measures, etc;

- **Data Transformation:** Transformation of data to the schema and needs of the destination system. This could include standardization of data, structure and restructure.
- **Data De-Identification:** Removing or masking direct identifiers of users (names, numbers, address).

2.2.2.3 Load

The load phase is where the data that was previously transformed will be inserted into the new database. The goal of the loading stage is to create a consistent and efficient centralized repository of data for analytics, reporting, and another variety of purposes [17].

This is not as simple as a step as it seems to be, as it does not just involve dumping the data into the new database. It involves identifying the target database in advance, defining the loading strategy and implement data validation mechanisms to ensure the integrity, completeness, and accuracy of the loaded data.

ETL is a crucial process for healthcare organizations as it enables them to collect, organize, and analyse vast amounts of patient information, such as EHR, clinical systems, and medical devices and consolidating all that data into a consistent, unified place [28].

One key benefit of ETL is its ability to support systems needing constant updates on their data [6]. This enables the implementation of services that can leverage the strengths of real-time streaming and ETL. By combining both processes, healthcare organizations can achieve efficient data management, real-time insights and thorough analysis, leading to improvement in the patient treatment.

2.2.3 Power BI

EHR have revolutionized how patient information is stored, accessed, and analysed, providing a comprehensive digital repository for all aspects of patient care. However, the volume and complexity of EHR data present significant challenges for healthcare providers. This is where advanced analytics tools like Power BI come into play.

Power BI is a robust business analytics tool developed by Microsoft that is responsible for enabling users to visualize data, make all kinds of different and interactive reports, and gain actionable insights [63]. As a cloud-based platform, Power BI connects to a variety of data sources and turns raw data in relevant information. Among other things, PowerBI have:

- **User-Friendly Interface:** designed to be intuitive to everyone, allows users to create and customize reports and dashboards without extensive technical knowledge [63].
- **Data Connectivity and Integration:** can connect to a wide range of data sources, including SQL databases, Excel spreadsheets, and cloud services, making it versatile for use across various industries [64]. This connectivity facilitates the integration

of disparate data types, enabling comprehensive analyses that consider all relevant data;

- **Real-Time Data Analysis and Monitoring:** supports live data updates, allowing users to monitor and analyse live data feeds directly within their reports and dashboards;
- **Advanced Analytics and Visualizations:** provides a variety of data visualization options, like graphs, tables, charts, etc.

As said by [63], mastering Power BI can improve an organization's capacity to leverage data power and promote well-informed decision-making. In healthcare particularly, by using Power BI, healthcare providers can effectively visualize and analyse large datasets of patient information, ensuring adherence to clinical guidelines and improving overall healthcare delivery.

2.3 Data Analytics Databases

Data analytics databases are specialized read-only database management systems designed to fit the specific needs of analytical workloads. A large amount of organizations produce massive volumes of data every hour [29], so, the storage, management and analysis of this data is essential to maintain the healthy workflow of these systems. Unlike operational databases, which prioritize data consistency [27] and transactional integrity [9] to support real-time operations, data analytics databases prioritize efficient querying [29], aggregation [76], and manipulation of large datasets [54] and are optimized for heavy workloads, complex queries, and ad-hoc analysis [51]. Introducing analytical workloads to operational databases can degrade their performance and disrupt critical business processes.

Data analytics databases are designed to go beyond simple data retrieval, focusing on deep data analysis and aggregation [18] and can come in various forms, including relational databases, NoSQL databases, columnar databases, and in-memory databases. Each one gives distinct benefits and trade-offs in terms of scalability, flexibility, performance, and data consistency. Selecting the best database will always be based on factors, like the data nature and volume, the complexity of queries, and the specific requirements of the analytics workload.

Traditional database technology usually encounters a data processing issue since it must load and organize all of the data before it can be analysed. This approach is effective in most situations, but it becomes ineffective when quick, on-the-spot continuous processing of large amounts of data is required [45]. Data analytics databases, however, integrate with real-time streaming platforms to ingest, process, and analyse streaming data in real-time [58]. Thanks to this integration, organizations can gain actionable insights from streaming data sources providing businesses with immediate, actionable insights from

live data. Moreover, the ETL mechanism is also very important in these types of databases as they enable organizations to base their decisions on accurate, consistent, and current data, ensuring they are responding to the most relevant insights available [54].

2.3.1 Relational vs Non-Relational Databases

The ongoing advancement of technology has led to an exponential growth in the volume of data generated. To effectively manage this substantial influx, it is essential to select a suitable database [30] that stores that data in a consistent, stable, repeatable and quick manner [41]. According to [56], three fundamental requirements must be met by databases management systems. The stored data must be available when it is needed (**availability**), only accessible to authorized users (**confidentiality**), and modifiable solely by authorized personnel (**integrity**). Traditional relational database management systems (RDBMS) have been developed for years to meet these requirements.

Relational databases organize data in a structured manner, using tables that consist of rows and columns, where each row corresponds to a record (e.g., a patient), and each column holds a specific attribute (e.g., name, address) [70]. These tables have defined relationships with each other, often through unique identifiers, allowing for an efficient connection and analysis of related data points, making it a predominant choice for storing data over hierarchical model or network model [41].

The dependability of relational databases is guaranteed by the ACID (Atomicity, Consistency, Isolation, Durability) properties. According to [41],:

- **Atomicity** ensures that if any part of a transaction fails, the entire transaction is considered incomplete;
- **Consistency** certify that the database remains stable at all times.
- **Isolation** guarantees the separation between concurrent transactions.
- **Durability** ensures that, even in the presence of an error or crash, a committed transaction remains in the same state.

MySQL and Oracle are the two most widely utilized relational databases. MySQL (Section 2.3.4) is more popular with the websites due to its lightweight system and quickness and Oracle is majorly used in case of large database requirements [19].

Relational databases, though, face challenges when dealing with modern data demands. Their fixed schema struggles to adapt to evolving data models [70], leading to difficulties in managing and updating information [30]. Additionally, performance issues also occur in dealing with huge amounts of unstructured data including multimedia and social media information [56].

Non-Relational databases, also known as NoSQL databases, however, tackle this issues with success, offering scalability, performance, and storage capabilities necessary for modern data environments [56] [69].

NoSQL, which means "Not Only SQL", suggests that while developing a software solution or product, many storage mechanisms can be used [56]. NoSQL databases can be categorized into four types:

- **Key-Value databases:** the simplest type of DB to use, as it is just a list of key-value pairs stored within an entity [70].
- **Document databases:** store hierarchical data structures directly, allowing documents to be referenced via unique keys [56].
- **Column Oriented Databases:** data from a single column is stored in each storage block [23].
- **Graph Databases:** store entities and relationships between these entities [30].

Additionally, NoSQL excel at managing unstructured data such as emails, multimedia files and word processing documents [44].

Selecting an appropriate database for a particular application involves a comprehensive assessment of the application's demands and potential challenges, as well as a careful evaluation of the business objectives and priorities [56]. Here is a table (Table 2.3.1) with the characteristics of each type of database, some of which have already been mentioned in this chapter, taken from the following articles [41] [13] [23] [56] [19] [70] [30].

2.3.2 MongoDB

MongoDB is a widely-used, open-source database that follows a document-oriented model and is recognized as one of the leading NoSQL databases. It utilizes a binary representation of JSON-like documents and embraces dynamic schemas, setting it apart from conventional relational databases [79].

MongoDB, being an example of a NoSQL database, is going to inherit all of the advantages that were highlighted in the previous chapter, such as the performance, flexibility and scalability. Therefore, the real question becomes, why use MongoDB and how does it help in the particular field of healthcare?

Firstly, MongoDB is classified as a document store, as opposed to other types like key-value stores, column stores, or graph databases, which were talked about in the previous chapter. This classification highlights MongoDB's strength in handling complex, nested data structures, providing the flexibility necessary by allowing healthcare providers to store various data types, including structured data (e.g., patient demographics, lab results), semi-structured data (e.g., JSON documents for patient encounters), and unstructured data (e.g., clinical notes and medical images), all within a single database. This flexibility is crucial for EHR systems, enabling the data model to evolve as new data types are introduced, in contrast to traditional relational databases that require predefined schemas [15]. MongoDB also excels in scalability by supporting horizontal scaling, allowing

Table 2.1: Relational vs Non-Relational

Feature	Relational	Non-Relational
Performance	Low	High
Scalability	Vertical scalability requires more powerful hardware for scaling.	Horizontal scalability allows for straightforward expansion by integrating additional nodes.
Query Language	Employs SQL for data retrieval and manipulation tasks	Uses their own query languages
Schema	Predefined schema	Has the capability to support multiple data formats
Development	Demands a greater investment in development resources.	Simpler to implement and demands fewer resources
Complexity	Medium	Low
Security	High	Less secure
Flexibility	Low flexibility	High flexibility
Transactional properties	ACID (Atomicity, Consistency, Isolation and Durability) properties	BASE (Basically Available, Soft State, Eventual Consistency) properties
Complex queries	Used	Not Used
Use cases	Particularly effective for structured data that has defined relationships, necessitating intricate queries and ensuring high data integrity	Perfectly suited for handling unstructured data, real-time processing, big data analysis and dynamic data needs

healthcare organizations to expand the database across multiple servers as the volume of patient data grows [10]. With sharding, MongoDB distributes data across machines, balancing the load and maintaining high performance and availability, essential for organizations managing large datasets [42]. Performance is another key advantage, as MongoDB's efficient indexing and optimized query engine enable quick retrieval and analysis of data, making it ideal for real-time applications like clinical decision support systems and compliance monitoring tools [20].

As highlighted by the Figure 2.3, in MongoDB, data is stored in collections and documents rather than tables and rows, as in relational databases. At the top, obviously, we have the databases, which are containers for multiple collections [31]. Unlike tables, collections do not adhere to a predefined schema, allowing them to keep data stored in the form of BSON documents (binary-encoded JSON-like structures). Documents are the basic unit of storage in MongoDB, representing a single piece of data [31] and each collection can contain multiple documents. Each document includes an ID field that functions as its primary key, along with a collection of key-value pairs. Moreover, these documents feature a dynamic schema, meaning that documents within the same collection are not required to share identical fields or structures. Consequently, common fields within the documents may store varying data types [20]. This structure is particularly beneficial for storing EHR data, which can vary significantly from one patient to another.

It is also important to note that MongoDB's robust security features and support for data encryption and access controls make it suitable for storing sensitive healthcare data in compliance with regulations such as HIPAA(Health Insurance Portability and Accountability Act). This compliance is critical for protecting patient privacy and ensuring that healthcare organizations meet legal and ethical standards.

2.3.3 DocumentDB

A document serves as a record that encapsulates data about a specific object along with its associated metadata, using key-value pairs to organize data in various types and structures, such as strings, numbers, dates, arrays, or objects. This stored data is structured as documents in XML, JSON (Figure 2.2), or BSON formats, providing flexibility in schema management and logical grouping of information [19].

DocumentDB is a fully managed NoSQL database service offered by Amazon Web Services (AWS) designed to manage, store and organize data as a collection of documents [50] [44]. DocumentDB is the AWS distribution that is built on top of MongoDB, meaning it inherits many of the benefits and features of MongoDB, such as its flexible schema and intuitive data model. With DocumentDB data is more logically grouped together [23] permitting users to incorporate an unlimited number of fields with diverse lengths into a document [44]. DocumentDB also takes advantage of connectivity with other AWS services, such as AWS Lambda, a serverless computing service that allows code execution in response to specific events, and AWS Glue, a fully managed data

integration service that simplifies data preparation and movement for analytics. It integrates seamlessly with AWS's own security features, including Identity and Access Management (IAM) and encryption services, to provide a secure and scalable environment.

Most of the advantages of DocumentDB have already been mentioned in Section 2.3.1, but it is important to mention that, in addition to their intuitive structure, DocumentDB stands out in its ability to handle unstructured, semi-structured, and structured data [16]. DocumentDB is particularly good at handling large-scale collections of literal documents [16], such as text or XML documents, as well as "sparse" data that might otherwise necessitate the extensive use of "null" values [50].

DocumentDB also supports change streams, which are a feature of MongoDB. Change streams allow applications to listen for changes that occur in the database in real-time. These change streams provide continuous feeds of data, delivering notifications whenever documents are inserted, updated, or deleted, making them valuable tools for real-time applications and data workflows, ensuring users stay updated on modifications within their database.

```
{
  "_id": "1234567890",
  "title": "Sample Document",
  "content": "This is a sample document stored in DocumentDB.",
  "author": {
    "name": "John Doe",
    "email": "john.doe@example.com"
  },
  "tags": ["sample", "document", "DocumentDB"],
  "created_at": "2024-02-07T12:00:00Z",
  "updated_at": "2024-02-07T14:30:00Z"
}
```

Figure 2.2: Document example

2.3.4 MySQL

MySQL stands out as one of the most popular open-source relational database management systems (RDBMS), owned by Oracle Corporation. It has become a foundational tool for data storage and retrieval across a diverse range of applications, spanning from small web applications to complex enterprise systems [74]. Distinguished for its scalability, dependability, and impressive performance, MySQL is often the go-to choice for numerous developers and organizations.

As opposed to MongoDB, MySQL is a relational database management system that stores data in tables, where each table consists of rows (records) and columns (fields) (Figure 2.3). It uses SQL (Structured Query Language) for database management and

manipulation, allowing users to create, modify, and query data efficiently. This facilitates effective storage, retrieval, and management of structured data [8].

MySQL open-source nature allows for cost-effective implementation, and it is highly scalable, capable of handling everything from small applications to large-scale systems with millions of queries. The system is known for its speed and efficiency, which is crucial when using applications needing quick accessibility to large datasets, such as healthcare systems. Such speed is also necessary for applications that require constant accessibility of data such as the web applications or graphical interfaces [78].

MySQL also supports complex queries, join and aggregation so that data could be retrieved, sometimes from different tables, and used in various manners in order to address the demands coming from dynamic user interfaces [8]. This helps in generating data for charts, dashboards or any web based environment which assure that the front end applications are always being fed with correct data. This point is particularly important in this particular thesis because it enables the handling of complex queries and integrate with analytics tools such as Power BI makes MySQL valuable for healthcare organizations that aim to assess and optimize clinical pathway compliance, identify treatment patterns, and track patient outcomes.

Another key advantage of MySQL is its compatibility with various platforms and operating systems, such as Windows, Linux, and macOS and its ability to integrate with a wide array of programming languages and frameworks also makes it versatile in different development environments [74].

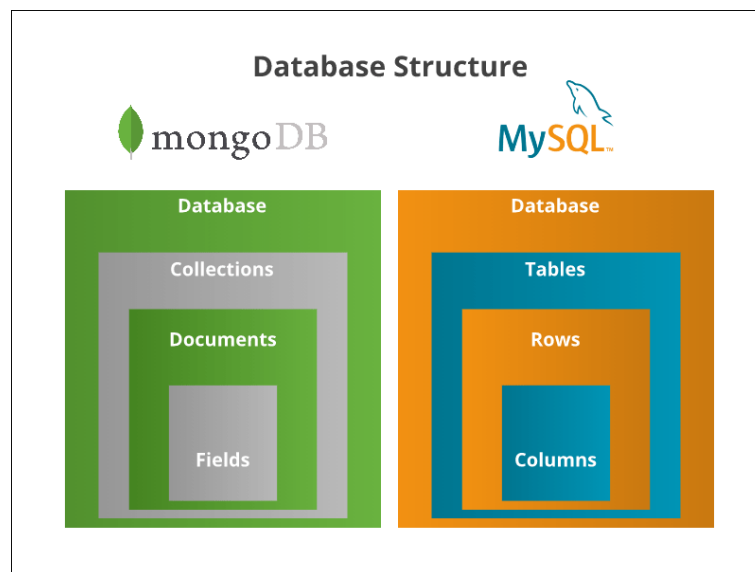


Figure 2.3: mongoDB vs MySQL

2.4 Clinical Data Repository

In order to be able to verify the compliance that exists between the CPs and the corresponding standardized CPWs, it is important to clarify how the information is delivered by the healthcare organizations, how it is transformed, mapped and inserted in the CDR in order for that compliance to take place. But first, what are the features that are essential in order to achieve those actions?

2.4.1 Definition

The delivery of healthcare is a very complex process that requires various professionals, resources, and functions. As a result, there is a significant amount of dispersed and different type of data [32] and effectively organizing these large volumes of clinical data is essential. Therefore, the notion of CDR becomes increasingly important as the demand for efficient data storage continues to rise.

The CDR is a centralized and secure platform for storing, managing and organizing patient information related to health collected from different sources in the healthcare environment [55]. This repository usually embraces extensive clinical information, for example, EHRs, medical imaging reports, laboratory results, medication histories and other relevant patient data. By aggregating all this information, CDRs not only provide a comprehensive and unified view of a patient's healthcare journey, but also improves healthcare delivery, facilitate clinical decision-making, and support medical research and analysis.

These repositories play a significant role in achieving interoperability across healthcare systems and ensuring authorized professionals have access to correct and timely patient information. CDR has a variety of components, designed to create an integrated system that includes various types of clinical data. Some of those components are:

- **Data Storage:** CDRs implement reliable data storage facilities that can deal with large amounts of structured and unstructured information.
- **Data Integration Engines:** These engines guarantee smooth integration of data from various sources through consistent and accurate storage in a repository.
- **Security Measures:** Healthcare data used by CDRs is highly confidential and, therefore, safeguarded using several security practices including access controls, as well as encryption to protect patient information and comply with healthcare regulations.

These components allow the CDR to enable a patient-centric approach to healthcare, allowing providers to deliver personalized interventions based on an individual's unique medical history, provide a safer way to handle patient data and improve efficiency, by removing the need of manual CPs and paperwork.

2.4.2 Healthcare Data Standards

Healthcare services worldwide have undergone significant transformation over time, with a growing demand for adherence to elevated quality and optimized standards. This shift has made the examination of healthcare processes an important area of study, resulting in the creation of various methods and tools designed to improve the effectiveness of these operations.

Healthcare data standards are a set of agreed-upon guidelines and specifications that define how medical information is electronically exchanged, stored, and interpreted. These standards have critical functions in ensuring that health information is properly and effectively delivered between different healthcare systems and entities, helping to advance patient care, research and public health initiatives. They form the backbone of clinical data exchange, system integration, and clinical information systems [40]. HL7v2 and FHIR are some of those standards and together they contribute to the interoperability of healthcare systems, facilitating the smooth and efficient exchange of patient data across diverse healthcare environments.

2.4.2.1 Role of Interoperability in Standardization

The standardization of CPs is important to be able to efficiently increase the care delivery, reduce risk of errors and reduce costs. Nevertheless, this standardization is not easy to reach, as medical information systems are always in a constant process of evolution, which makes every single one of them a heterogeneous, distributed system with high complexity [33]. Additionally, sharing of health data between independent organizations brings major concerns, such as the different representation of data, communication barriers, confidentiality and patient privacy. Thus, it is essential to have an integrated environment commonly understood by the providers of healthcare [12] to cope with this concerns.

Healthcare information must be made widely accessible and simple to understand for both humans and machines to arrive at a more clear understanding of the healthcare system [4]. With this comes the importance of interoperability, to enable the connection, integration, and effective utilization of healthcare data across systems, organizations, and users to make sure that various types of data, when needed, can interact with each other [33].

Healthcare interoperability refers to the seamless exchange, interpretation, and utilization of data among various health information systems, applications, and devices. Basically, it ensures that systems that are often developed independently, can work together. Here, healthcare data standards plays a very important role, as they provide a common language for representing and exchanging clinical information, enabling healthcare providers to access and share patient data without problems, regardless of the specific system or software they use. CDRs are enablers of interoperability as they serve as mediating repositories that aggregate and manage heterogeneous clinical data from different sources.

Hospitals aiming for effective interoperability should first focus on digitalizing their processes. For this, building an IT framework that supports the deployment and management of these systems is essential. However, this is not only what is needed, as it is equally important to manage relationships with suppliers and ensure healthcare professionals are encouraged to adopt and use the technology [25]. As electronic medical information systems improve and technologies architectures that are capable of interoperability become even more present in the healthcare industry, resources to solve medical issues become more effective, information becomes more reliable, and the clinical decision-making process becomes easier [71].

2.4.2.2 Health Level 7 v2

For a long time, interoperability was achieved through paper-based communication and healthcare professionals had to rely on fax machines and paper documents in order to share essential patient data. With the evolution of technology, suddenly it became apparent that all this work could be done through communications between computer systems. However, the challenge remained to ensure that this data transfer was secure, standardized, and protected in terms of privacy [11]. Due to the variety of software applications used across healthcare systems, it was crucial to establish a unified protocol for seamless data sharing [71]. Because of this, the Health Level Seven (HL7) was created.

HL7 is a set of international standards created by Health Level Seven International for facilitating the exchange, integration, sharing and retrieval of electronic health records between healthcare systems. This means that the HL7 provides the standard guidelines (language, structure and data types) on how the data should be shared and retrieved from one system to another.

HL7v2 is one of the versions of HL7 and it is one of the foundational standards for data exchange in helthcare settings. It is a widely adopted and common standard for hospitals to send their data due to its in-patient settings. This version uses messages to transfer information between systems, covering a wide range of healthcare events, such as patient admissions, test results, and treatment orders. These messages have a variety of fields, typically in a pipe-delimited format (HL7-ER7), each containing specific information about a healthcare event or transaction. These fields are divided into many segments and again in message groups. It is possible to include a variety of types of health data in the HL7v2 messages, including demographics, medical history, test results and medication orders. Figure 2.4 is an example of a HL7v2 message.

```
MSH|^~\&|GHH LAB|ELAB-3|GHH OE|BLDG4|200202150930||ORU^R01|CNTRL-3456|P|2.4<cr>
PID|||555-44-4444||EVERYWOMAN^EVE^E^^^L|JONES|19620320|F|||153 FERNWOOD DR.^
^STATESVILLE^OH^35292|||(206) 3345232|(206) 752-121|||AC555444444|67-A4335^OH^20030520<cr>
OBR|1|845439^GHH OE|1045813^GHH LAB|1554-5^GLUCOSE|||200202150730|||
555-55-5555^PRIMARY^PATRICIA P^^^^MD^^|F|||444-44-4444^HIPPOCRATES^HOWARD H^^^^MD<cr>
OBX|1|SN|1554-5^GLUCOSE^POST 12H CFST:MCNC:PT:SER/PLAS:QN|^182|mg/dl|70_105|H||F<cr>
```

Figure 2.4: HL7v2 message example

Its structure and complexity can represent a challenge in some cases, but the HL7v2 is still a very effective tool for communicating clinical data between healthcare organizations as it supports the majority of the common interfaces used in the healthcare industry globally. Not only that, but since it has been around for a long time, the search, analysis and re-use of this standard has become very efficient for the ones that use it [53].

However, while HL7v2 has been instrumental in enabling data exchange in healthcare, as it can be verified by the complex representation in Figure 2.4, it has also been criticized for its complexity. This has led to the development of newer standards, like FHIR, which offer a more lightweight and modern approach to healthcare data exchange.

2.4.2.3 Fast Healthcare Interoperability Resources

In 2012, in response to the growing need for faster, more efficient, and user-friendly ways to transmit health data, the development of FHIR began. Fast Healthcare Interoperability Resources (FHIR) is a modern and web-friendly healthcare interoperability standard developed by Health Level Seven International (HL7). It integrates the strengths of earlier HL7 versions and CDA documents [53] while, at the same time, leveraging modern web technologies to achieve a more secure, cost-effective and highly flexible interoperability standard [71] [43].

At its core, FHIR is built around “Resources” that represent key clinical concepts. These resources, such as Patient, Practitioner, Medication, and Observation, can be used individually or combined to form more complex structures.

Applications or systems developed using the FHIR standard can be based on one of the following paradigms: RESTful interfaces, Messaging, Documents, Services, Persistence/Databases and Subscriptions Framework. In terms of the RESTful interfaces, it highlights one of the main goals of FHIR, which was to make a standard that would be very easy to understand and use across the developers community. The RESTful interfaces in FHIR are defined by several notable features that enhance its usability, particularly for the software developers. One of them is the facilitated integration through the use of HTTP methods, ensuring accessibility for developers. Each resource is linked to a directory-structured URI, allowing for easy identification and navigation, and resource data is transferred in formats such as XML (Figure 2.5), JSON or RDF based. This is a very easy to understand representation of the advantage of FHIR in comparison to prior standards, as the messages becomes much more clear and user-friendly than the ones represented by, for example, HL7v2 [43] [75].

In addition to its flexibility, FHIR also prioritizes security and privacy. It incorporates modern standards for encryption and authentication, ensuring that sensitive patient information is protected.

Besides this, FHIR helps streamline workflows by enabling healthcare providers to retrieve only the most relevant clinical information, such as allergies when prescribing medications, enhancing both privacy and efficiency [49].

```

<DiagnosticOrder>
  <id value="845439"/>
  <text>
    <status value="generated"/>
    <div xmlns="http://www.w3.org/1999/xhtml">Glucose post 12h order for Eve Everywoman, completed on 2002-02-15T07:30:00.</div>
  </text>
  <subject>
    <reference value="Patient/555-44-4444"/>
    <display value="Eve E. Everywoman"/>
  </subject>
  <orderer>
    <reference value="Practitioner/555-55-5555"/>
    <display value="Patricia Primary MD"/>
  </orderer>
  <identifier>
    <system value="http://ghh.org/oe/placerorder"/>
    <value value="845439"/>
  </identifier>
  <status value="completed"/>
  <event>
    <status value="completed"/>
    <dateTime value="2002-02-15T07:30:00"/>
  </event>
  <item>
    <code>
      <coding>
        <system value="http://loinc.org"/>
        <code value="1554-5"/>
        <display value="GLUCOSE^POST 12H CFST:MCNC:PT:SER/PLAS:QN"/>
      </coding>
    </code>
    <status value="completed"/>
  </item>
</DiagnosticOrder>

```

Figure 2.5: FHIR message example

As a carrier of innovation, the FHIR has played a very important role in facilitating data integration and enabling better patient care while also increasing the efficiency [71].

2.4.3 From HL7v2 to CDR

Despite the fact that HL7 FHIR is recognized as the current best available standard to achieve interoperability between healthcare systems, not all institutions have adhered to this new "trend" and still use the main method of HL7v2. Therefore, in order to fill the CDR with the most up-to-date standards, such as the FHIR Resources, it needs to happen a transformation in this data received from those hospitals. These transformations are highlighted in the following diagram (Figure 2.6).

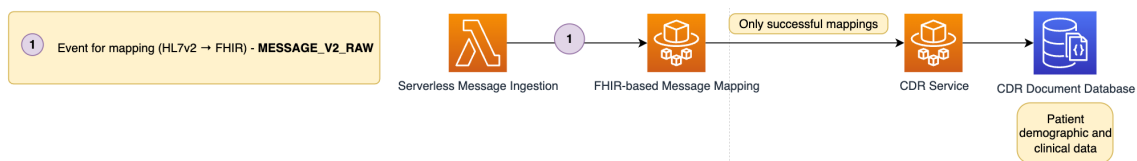


Figure 2.6: FHIR Mapping

As seen by the diagram, the process of transitioning from HL7v2 messages to FHIR resources and storing the data in a CDR involves several key steps. Below is a sequential guide outlining these steps:

1. **Serverless Message Ingestion:** HL7v2 messages are received from various sources, such as hospitals, laboratories and/or pharmacies.
2. **HL7v2 Message Parsing:** decompose the HL7v2 message into its components. This includes obtaining the data fields from the message and putting them into

a structured form.

3. **FHIR-based Message Mapping:** As soon as the HL7v2 message is parsed, it gets mapped to the relevant FHIR resources. First, the appropriate FHIR resource type for every data element is found and then its values are mapped accordingly.
4. **FHIR Resource Creation:** generating the specific JSON or XML representation of the FHIR resource, including the resource type, resource ID, and data elements.
5. **CDR Service:** The transformed FHIR resources undergo validation to ensure compliance with the FHIR standard and to identify any data integrity issues.
6. **FHIR Resource Storage:** The FHIR resources that are created will then be stored in a CDR.

This transformation allows for an easier and more clear way of representing the CPs that are provided by the healthcare organizations. Without this, the compliance checking of the CPs to the standardized CPWs would probably have to be done by manual analysis, which would not be an efficient and scalable way to make that compliance. Furthermore, that transformation also allows for a real-time compliance checking.

SOLUTION

In this chapter I will explain the solution applied to develop this thesis and how the different components mentioned in the state of the art were used to arrive at the most efficient solution that will give me the best possible results. This chapter is divided into 2 topics. The first concerns the methodology of the solution and the initial work that was needed in order to implement the best solution possible. The second section concerns the solution itself and is divided into several sub-topics. A small diagram of the solution that was implemented will be presented and then, for each block of the diagram, a sub-topic was created where it will be explained what was done and the impact of the work carried out in this step on the final result.

3.1 Methodology

As mentioned earlier, the whole process of comparing compliance would be perfect if all there was to do was take a clinical file from the hospital, go through each step, check which ones are compliant and which ones are not and draw the appropriate conclusions. However, this compliance mechanism is not so simple and there is a whole process that has to be carried out in order to arrive at the results.

The first step was to define the ideal pathway for the compliance assessment. It was necessary to have a pathway that had the following attributes:

- divided into phases to be able to see the evolution of compliance throughout the treatment phases;
- it had patient sessions being currently created, so that they could be analysed in real time
- the patient sessions had a reasonable number of steps (more than 8 and less than 30).

After discussions with experts in the field, it was concluded that the ideal pathway would be that of colorectal cancer treatment at the Centro Hospitalar Universitário de Lisboa Central (CHULC).

The next step was to define the strategy regarding the best way to realise whether the patients' treatment were in compliance with the pathway developed. Here, the solution was to divide 2 important concepts, resources and patient sessions. Both concepts will be discussed over the next few sub-chapters, particularly the compliance-service, in order to better understand why they are important, but for now here is a brief explanation of what each means. A resource is a document that is captured by the change stream and carries clinical information about a specific patient. In this specific case, there are 3 types of resource, Encounter, Appointment and QuestionnaireResponse. These resources are particularly important in our implementation because, as will be explained in the CDC section, it is them, or rather their capture, that will start the development of the real-time solution. A patient session, as the name suggests, is the instantiation of a patient inside the healthcare system within a particular pathway. This includes the medical treatment that doctors need to perform, along with all the steps and procedures involved in that treatment.

3.2 Solution

The diagram provided (Figure 3.1) offers a visual representation of the services that were used in order to address the challenges outlined in this thesis.

3.2.1 Clinical Data Repository

The CDR is the element that served as basis for the solution. It was the centralized repository where all the patients' information and related data were stored. This information in the CDR was stored under the FHIR standard, using the better and most reliable healthcare data standard and ensuring that clinical data was structured in a uniform and interoperable manner. DocumentDB was used to store those FHIR Resources due to its seamless integration with the AWS ecosystem, which provided an additional layer of encryption for data, enhancing security. By using a NoSQL Database, the CDR ensures flexibility in handling various types of clinical data, from unstructured clinical notes and medical images, to structured patient records, and scalability and robustness, by handling the large volumes of data generated in healthcare settings while maintaining high performance and reliability.

3.2.2 Change Data Capture

Change Data Capture (CDC) is the first implemented component of the solution. CDC is a very important tool, designed to capture and track changes occurring within the CDR in real-time. This is where the real-time streaming process mentioned in Section 2.2.1 comes into play. This capability ensures that any updates to the CDR are immediately reflected in downstream processes. CDC's role is particularly important for maintaining up-to-date compliance with clinical processes in the analytics-service, as it enables real-time data

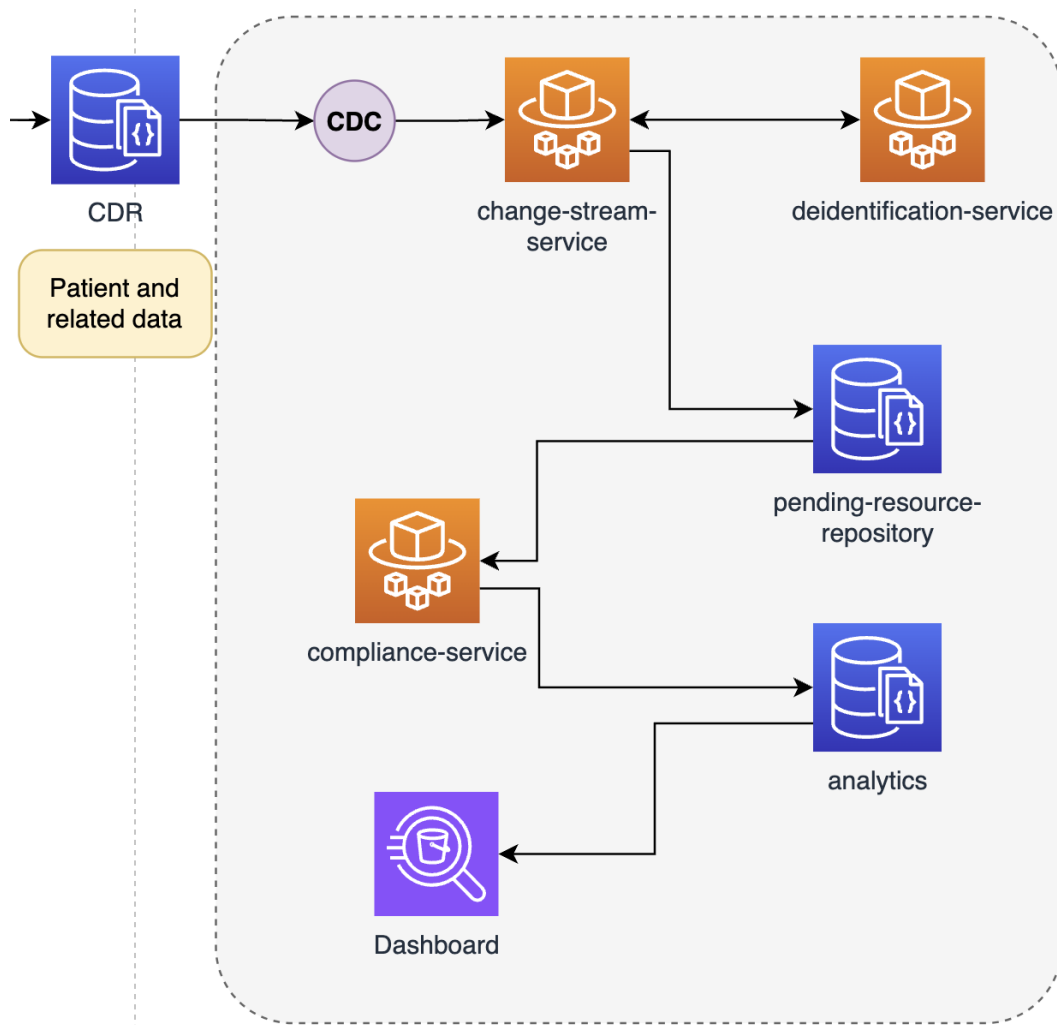


Figure 3.1: Solution

synchronization. The real-time streaming process is powered by CDC to ensure that all compliance checks are performed with the most current clinical data. The moment a change is detected in the CDR, CDC triggers an automatic update to the compliance service, ensuring that analytics and decision-making processes are always based on the most accurate and timely information. This dynamic process helps maintain the integrity of real-time streaming, making it both efficient and highly responsive to changes in clinical workflows.

CDC is essentially the starting point of the ETL (Extract, Transform, Load) process (Section 2.2.2), focusing on the extraction phase. Each time the CDR is updated with a new Resource, CDC extracts the relevant data and immediately sends it to the change-stream-service, which prepares the data for further analysis.

In terms of pure implementation, firstly, I had to enable the DocumentDB change stream for all collections withing the database that I was working for. Then, DocumentDB, from AWS, together with a MongoDB API was used, particularly, by implementing one

of its methods, a change stream method, responsible for monitoring and capture any modifications made to the data stored within the collections selected. A database client compatible with MongoDB API is used to receive the Change Stream, allowing it to respond to these changes accordingly and store them in the pending-resources-repository.

3.2.3 Change-stream-service

The Change-stream-service serves as the entry point for processing and transforming data extracted from the CDR. This corresponds to the second phase of ETL, the Transformation phase. This phase transforms extracted data into a format that can be easily compared to CPWs. So, the Transformation phase ensures that all the information coming from the CDR is transformed/mapped and prepared to be stored in the pending-resources-repository and consequently analysed in the compliance-service.

As mentioned in the Section 2.4, one of the main challenges in working with data integration between different healthcare systems is reconciling the various data standards each system uses. UpHill, for example, has its internal data format, while Electronic Health Records (EHRs) often rely on HL7v2, which is then mapped to the FHIR standard. This discrepancy required a thorough process of aligning and standardizing data to ensure interoperability.

In the transformation phase, we addressed the complexity of the HL7 and FHIR standards, which include a variety of information to support diverse systems and use cases. Although comprehensive, this data often contains a lot of fields that are not directly relevant for compliance analysis. The resources used in this case were Encounter, Appointment, and QuestionnaireResponse and each of them stores key clinical concepts differently. Through the transformation process, these resources were rearranged into a unified format, selecting only the necessary information, such as resource type, timestamp, and patient identifier, to be able to efficiently analyse the compliance while maintaining the integrity of the clinical data.

This ensures that the data being processed and analysed are aligned with the specific objectives of the compliance-service. By standardizing and extracting the critical information of each resource, data quality is improved, processing overhead is reduced, and the overall process becomes more effective.

3.2.4 Deidentification-service

As noted in Section 2.2.2, data de-identification is the process of removing or altering personally identifiable information from a dataset, making individuals unidentifiable while preserving the value of the data for research, analysis, or sharing. This is a crucial step in healthcare data management, particularly within the analytics pipeline, where large volumes of sensitive patient data are processed.

In this solution, the Deidentification-Service tokenizes patient private information, such as name, address, phone number, using a Java method that is triggered every time

a new resource enters the CDR and is responsible for automatically hide any private patient data. This encoded data retains its usability within the system for compliance checks and analytics while ensuring that the original data is protected throughout the entire compliance process. By tokenizing sensitive information in this manner, the system ensures that the patient's private information remains confidential and secure while still allowing the healthcare data to be used effectively.

Besides this, it's important to remember that, as mentioned in the Section 2.3.3, this de-identified data is stored in DocumentDB, which was chosen for its high levels of security. DocumentDB provides encryption both at rest (ensuring that stored data remains secure in memory) and in transit (securing the data as it is queried and transmitted). This ensures the protection of sensitive patient information at every stage, while also guaranteeing system performance, scalability, and compliance with privacy regulations.

3.2.5 Pending-resource-repository

The Pending-resource-repository component in the diagram serves as the target system for the Load phase of the ETL process, where transformed data is loaded for subsequent analysis and evaluation. This is the "storage" where only the essential information about each patient's clinical journey is kept. This is the only information that is responsible to evaluate the compliance.

In the compliance-service, multiple external service calls were required to assess the compliance, which introduced various potential points of failure, particularly in a real-time solution. To address this challenge and ensure efficiency, this pending-resources-repository was created to act as a queue, managing the influx of data from the CDR. This queue temporarily stores incoming resources and processes them asynchronously every 30 seconds. This approach is crucial because the external services required by the compliance-service may not always be available in real time. Without the queue, any unavailability or failure in these services would result in the loss of data for the resources being processed. The queue ensures that resources are retained until they can be successfully processed, implementing a retry mechanism to handle failed attempts, which enhances fault tolerance and system reliability. Once the resources are processed, they are removed from the queue, clearing the way for new upcoming data. This solution not only makes the system more robust but also allows for a more resilient real-time compliance evaluation.

In this repository, as mentioned in the change-stream chapter, 4 essential fields were stored that were needed at different stages of the compliance process.

1. **resource id:** this helped identify the resource and understand whether it is compliant (it has been used to complete some patient-session action) or non-compliant (it has not been used to complete any action, so it is assumed to be extra);
2. **Type of resource:** helped us understand whether the resource was used to represent

an interaction between the patient and the doctor (Encounter), to represent an appointment that is planned for a certain day (Appointment), or to represent a patient's response to a questionnaire posed by the doctor (QuestionnaireResponse);

3. **Timestamp:** knowing the creation date of this resource, used later to evaluate the metrics of the pathway phases;
4. **Patient id:** identify the patient for whom this resource was created, so that, as we will see in the next chapter, we can retrieve the patients' session.

It is important to note that, in a scenario where these 4 components (CDC, change-stream-service, deidentification-service and pending-resource-repository) did not exist, the alternative implementation would involve a method in the compliance-service where you would go through the CDR, filter each CP individually, filter only the information that relates to the patient's journey and compare that information with the CPWs. This would make this process very inefficient, unscalable and with a high risk of errors, a very important aspect, not just when verifying compliance, but for any data processing system. This way it is guaranteed that there is a repository that makes it easy to access and analyse only the relevant data for compliance checking purposes and there is a mechanism capable of giving the best possible information to the compliance-service.

3.2.6 Compliance-service

The Compliance-Service is the central service of this thesis. It provides the code necessary for evaluating compliance with CPWs and assessing the effectiveness of healthcare processes.

The first step here was identifying the cases where non-compliance was present. In order to evaluate the causes of non-compliance, it is necessary to define the scenarios where that variations can happen. Those deviations are:

1. **Skipped activities:** Steps that should occur according to the CPW, but were not performed. For example, a cancer patient that needs a routine check-up to assess their condition, but due to time constraints, the clinician decides to skip the check-up.
2. **Inserted events:** These are activities that were performed by the clinician but were not part of the CPW's guidelines. For example, there are only 2 appointments scheduled for a certain patient and the clinician does 3 appointments with him. This is a particular important case, because it allows for healthcare organizations to identify scenarios in which they are doing more than they should and "wasting time" on tasks that, according to standardized guidelines, are not necessary. This is a clear indicator that an optimization can be done in the CPs done by the organization and, at the same time, in the management of the human resources of the healthcare organization.

Now how can we identify each scenario?

Skipped Activities

Before explaining how these missing steps are identified, it is important to clarify the concept of a patient-session. A patient-session is the instantiation of a patient's clinical journey in a CPW that aims to achieve specific clinical or therapeutic outcomes. In each session, a sequence of steps is defined, and each of these steps must be tracked to determine if it has been completed by the doctor. If a step is not performed, it is marked as "non-compliant," indicating a deviation from the treatment protocol. On the other hand, if the step is completed, it is marked as "compliant". In this case, this step carries the ID of the resource used to complete that step with it, which brings us to the second case of non-compliance.

Inserted events

Here, the search for events that have been inserted and were not supposed to is also very simple. If each step of a patient-session carries the id of the resource that completes it, this resource is automatically identified as compliant, because it was in fact used to complete a patient session step. If we reach the end of our compliance analysis and there are resources that have not been marked as compliant because they were not used to complete any step of that patient session, then that resource is marked as non-compliant. In other words, if a resource was used but did not contribute to completing any scheduled step, it is considered an inserted event and represents a deviation from the pathway.

After the entire compliance verification process was completed, both for the resources and for the steps present in the patient sessions, they were stored in a database so that they could be analysed to reach the results, and consequent insights into the hospital's processes.

3.2.7 Analytics

After the compliance analysis, it was time to think about the results and what would be the best way to gather as much information as possible to convey to the hospital so they can investigate what went wrong, what went well, and which aspects need improvement in order to make patient treatment even more efficient.

For this, I decided to put all the information that was analysed in the previous service into a sequential database, specifically a MySQL database, in order to take advantage of its vast benefits regarding query processing, which has already been detailed in the Section 2.3.4. I used the IDE DataGrip to create the required tables and subsequent views for each metric.

First, I started by creating the two main tables with the data collected from the compliance service. The first table was called 'compliance-resources' and pertains to all the resources that were collected by the change stream. This table contains the attributes that were in the pending-resources-repository, along with the new compliance attribute, a boolean that tells us whether this resource is compliant or not. The second table was

called 'compliance-steps' and refers to all the steps of all the patient sessions that were also processed in the compliance service. Here we have the following attributes:

- **Step ID** - primary identifier of the step;
- **Element** - the name of the specific element that this step corresponds to in the pathway;
- **Pathway** - name of the pathway this step belongs to;
- **Patient ID** - patient identifier;
- **Phase** - the name of the phase to which the specific step corresponds to in the pathway;
- **Timestamp** - date of the last modification of the step, or, if it was not modified, the creation date;
- **Compliance** - boolean indicating whether the step is compliant or not;
- **Index** - an integer that tells me the index of the step in the patient session.

The next step was to discover the best mathematical formula for calculating compliance. For this purpose, an intense research was conducted where it was decided to base it on the formula developed by [1], which was already mentioned in Section 2.1.3.

In order to make the next paragraph clear, it is important to remind that skipped activities are the steps that should occur according to the CPW, but are not being done and inserted events are the resources that are not being used to complete any step in the patient-session, meaning that they are non-compliant.

With that being said, skipped activities are identified with $A_s \subset A$, where A represents the set of all activities in the patient-session. All inserted events are identified as a set $E_i \subset E_c$, where E_i designates the number of events in the path travelled by patient c , $c \in C$, where C is the set of patients. This technique allows individual costs to skipped and inserted activities, taking into account the influence they may have on the CPW (to be defined by clinical professionals). Thus, k^s and k^i represent the cost functions for skipping and inserting activities, respectively. In this case, I simply assumed the same cost for both since the aim is just to pass on the information of which resources were incorrectly inserted and which steps were not completed in order to transmit this information. There is no logic as to which one has more impact. The function presented below defines fitness (f) as one minus the ratio between the total cost of having skipped/inserted activities and the total cost when considering that all events are inserted activities.

$$f = 1 - \frac{\sum_{a \in A_s} k^s(a) + \sum_{e \in E_i} k^i(\alpha(e))}{\sum_{e \in E_c} k^i(\alpha(e))}$$

in which $\alpha : E \rightarrow A$ represents a function relating each event of the case to an activity of the process.

In general, it is expected that a greater number of inserted or skipped activities will decrease the fitness value. In the worst possible case, all events are inserted, which is used to normalise the fitness metric. On the other hand, the maximum fitness value is reached when there are no activities of this nature.

Now it will be revealed and explained the metrics that were established in order to obtain the best possible results based on the solution that was implemented.

3.2.7.1 General Compliance

This metric provides an overview of the overall compliance to healthcare protocols and guidelines within the pathway. This metric is essential as it offers a snapshot of how well the healthcare services are aligned with the established pathway across all phases and activities. Higher general compliance results ensures that healthcare providers maintain a high level of service quality and patient safety, which is crucial for achieving better health outcomes. The query that was done was the following:

```
select pcev.'patient_id' AS 'patient_id',
((1 - ((pcev.'non_compliant_steps' +
        pcev.'extra_resources') /
        ((pcev.'non_compliant_steps' +
        pcev.'extra_resources') +
        pcev.'compliant_steps')) * 100) AS 'compliance_percentage'
from 'compliance'.'PathwayComplianceEquationValues' as pcev
where pathway_id = 'ccr_chlc_demoCA'
group by pcev.'patient_id'
```

The output of the query provides the compliance percentage. By grouping the results by patient id, the query ensures that the compliance score is calculated for each patient across the colorectal pathway, offering an individual assessment of compliance.

This compliance metric takes into account key elements like **compliant steps** (those followed correctly), **non-compliant steps** (missing steps that were not completed), and **extra resources** (any additional resource not part of the original plan). A 100% compliance score means all steps were completed as intended, while lower scores indicate inefficiencies or deviations that may require corrective action.

Assuming that the implementation of this query resulted in a view that was named "CompliancePercentagePerPatientSession", to then retrieve the average compliance of all the patient sessions is to just make an avg() function, like the following:

```
select AVG(cp.compliance_percentage)
from CompliancePercentagePerPatientSession as cp
```

3.2.7.2 Compliance by Action of the pathway

This metric is the more in-depth of all the ones presented in this section and it provides a detailed evaluation of how well each individual step/activity in a pathway is followed. This analysis is crucial to understand the exact points where the patient session may deviate from the established pathway.

In the colorectal cancer pathway, there are three types of events: Actions, which are the specific tasks performed by healthcare providers, such as administering treatments or scheduling appointments; Gateways, which serve as decision points that dictate the next steps based on patient evaluations; and Key Moments, which represent significant milestones in the patient journey, such as receiving a diagnosis or discussing treatment options. For the purpose of this analysis, I am focusing exclusively on the Action events. This is because Gateways are merely decision points and, by design, are always compliant, thus not providing meaningful insights for compliance assessment. Additionally, while Key Moments highlight important events in the patient experience, their compliance values are not relevant for this metric, as they do not require direct input from healthcare providers.

```
select cs.'element_id'                                AS 'elementId',
       sum((case
           when (cs.'compliance' = true) then 1
           else 0 end))                               AS 'true count',
       count(0)                                       AS 'total count',
       ((sum((case
           when (cs.'compliance' = true) then 1 else 0 end))
          / count(0)) * 100)                         AS 'percentage'
from 'compliance'.'compliance_steps' as cs
where ((cs.'pathway_id' = 'ccr_chlc_demoCA'))
and ((cs.'element_id' like 'ACTION%'))
group by cs.'element_id'
```

This query retrieves the elementId (representing each individual step in the pathway) and calculates the total number of compliant cases (**true count**) in all patient sessions, as well as the overall number of cases (**total count**). The compliance percentage is then determined by dividing the "true count" by the "total count" and multiplying by 100 to express it as a percentage. Finally, grouping the data by elementId allows the query to provide a breakdown of compliance for each specific step in the pathway.

By measuring compliance at this level, healthcare providers can ensure that every step, whether it is a diagnostic test, treatment, or follow-up, is being carried out as expected. Tracking compliance for each element in the pathway helps identify areas where improvements can be made and, ultimately, it leads to a higher quality of patient treatment and better health outcomes by ensuring that critical steps in the clinical process are not overlooked.

3.2.7.3 Compliance by Phase of the pathway

This metric breaks down the compliance at each distinct phase of a patient's journey through the CPW. Phases in a CPW are distinct stages of a patient's treatment, each outlining specific steps or interventions to ensure standardized and efficient care delivery. This detailed perspective is important because each phase, from initial diagnosis to treatment and follow-up, may have specific requirements that must be met to ensure effective patient outcomes. By analysing compliance at this level, healthcare providers can pinpoint specific phases where compliance may be lower.

```
select pcps.'phase_id'                                AS 'phase_id',  
      avg(pcps.'compliance_percentage') AS 'compliance_percentage'  
from 'compliance'.'PhaseCompliancePerSession' as pcps  
where pathway_id='ccr_chlc_demoCA'  
group by pcps.'phase_id'
```

This query retrieves the phase id and calculates the average compliance for each phase by selecting data from the "PhaseCompliancePerSession" table (as the name says, in here it is calculated the phase compliance for each patient session). The results are grouped by phase id, allowing the query to show a distinct compliance value for each stage of the pathway.

This query's advantage is, obviously, its ability to highlight which phases of the pathway have high or low compliance. If the compliance percentage in a certain phase is significantly lower than in others, it may suggest gaps in the treatment of patients that need to be addressed. On the other hand, high compliance percentages reflect well-executed phases, indicating that the clinical team is following the prescribed steps effectively.

3.2.7.4 Compliance per Week

This metric examines how the compliance fluctuated per week since the beginning of the data processing. However, it is important to note that this is a cumulative calculation, meaning that the compliance value for each week is not isolated to that specific week alone. Instead, each week's compliance is calculated based on everything that has been completed up until that point, accumulating all previous data. Monitoring compliance on a cumulative, week-by-week basis is critical for identifying trends, patterns, or sudden deviations that may indicate underlying issues. Weekly data can help healthcare administrators respond promptly to shifts in compliance, whether they are related to staffing, operational changes, or external factors.

```
select wpcv.'week_number'                            AS 'week_number',  
      wpcv.'week_start_date'                        AS 'week_start_date',  
      avg(wpcv.'average_compliance') AS 'average_weekly_compliance'  
from 'compliance'.'WeeklyPatientComplianceValues' as wpcv  
where pathway_id='ccr_chlc_demoCA'
```

```
group by wpcv. 'week_number' ,  
         wpcv. 'week_start_date '  
order by wpcv. 'week_number '
```

This query selects the index of the week, the start date of the week, which will be the date shown in the output, and calculates the average compliance for each week by averaging the compliance values from the `WeeklyPatientComplianceValues` (this view calculates the compliance for each patient-session per week). Grouping by the number of the week and the date ensures that the data is organized by the week, and ordering by the first one helps display the results chronologically. This query provides healthcare administrators with a clear, time-based view of weekly compliance.

The advantage of this approach is that it provides a clear and time-based view of compliance, helping healthcare organizations to detect patterns such as drops in compliance during certain periods (e.g., holidays, high patient influx) or consistent underperformance in certain weeks. By having this data readily available, healthcare providers can implement targeted strategies to improve adherence to pathways, ensuring that patient care remains consistent and of high quality throughout the year.

3.2.8 Dashboard

Finally, after completing the entire ETL process to prepare the data for analysis, doing the comprehensive compliance check and implementing specific queries and come up with metrics to reflect the key health indicators and performance metrics that doctors would need to make informed decisions, it was time to present this information in a format that was intuitive and easy to understand. To achieve this, I decided to create a dashboard, providing a streamlined way for doctors to access, analyse, and act upon the data.

The best way to present this information was through the use of analytical graphics, charts, and other visual elements that effectively transformed raw data into an accessible and visually engaging format. Power BI was chosen as the ideal tool for creating the dashboard due to several key advantages. As one of the most widely used platforms for data analytics, Power BI is well-known for its user-friendly interface, which makes it easy for users with varying levels of technical expertise to interact with data. Additionally, its robust library of graphs, tables, and charts offers diverse visualization options to suit different types of metrics and insights. These strengths made it the perfect tool for transforming data from MySQL tables and views into meaningful visual representations.

To accomplish this, Power BI was chosen as the ideal tool for creating the dashboard. As presented in Section 2.2.3, PowerBI's versatility, powerful data visualization and user-friendly interface made it very easy to transform the data from the tables and views of the MySQL databases to the graphics representation. Therefore the first thing to do was to create a workspace in PowerBI in order to store my information (we can see that workspace in Figure 3.2). Then, once all the information in the MySQL tables had been processed, it was exported to an Excel file or a CSV file and then uploaded to a report created on the

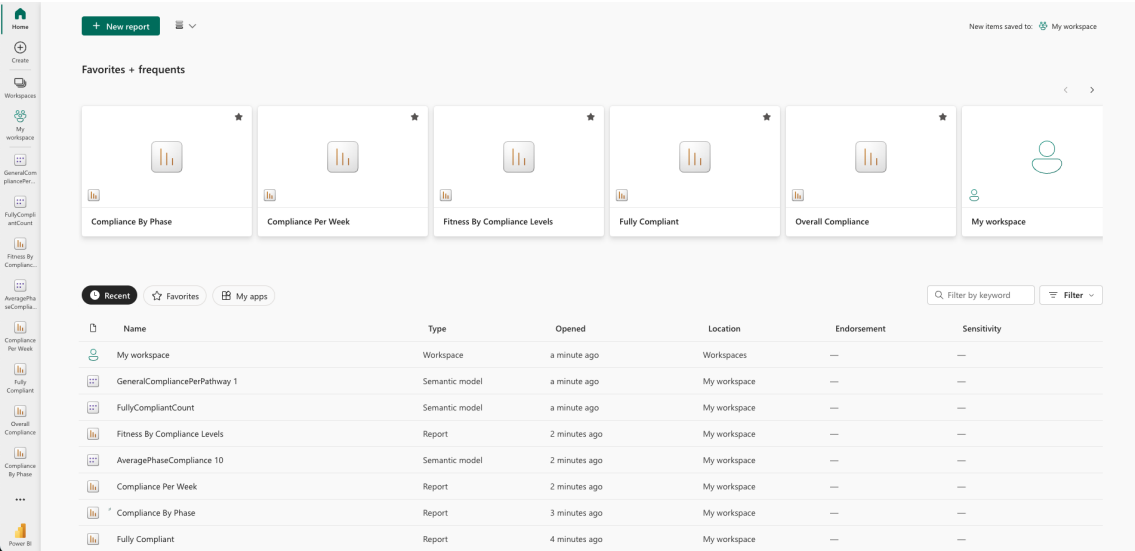


Figure 3.2: PowerBI workspace

workspace. Finally, it was just a matter of choosing the best visualisation for each type of metric.

RESULTS

4.1 Data processing

During the analysis of compliance within the clinical pathways, I observed a critical distinction: some steps within the patient sessions were configured with interoperability, while others were not. By interoperability configuration, I refer to the process where the completion status of certain steps is automatically updated via HL7v2 messages from hospital systems. However, for steps not configured with this integration, healthcare professionals (HCPs) were required to manually mark them as completed within the UpHill software. This created a scenario where HCPs had to duplicate their efforts—first recording their actions in their day-to-day EHRs, and then manually updating the UpHill system due to the lack of automatic information flow. The impact of this manual process became evident in the compliance data. It was clear that HCPs, understandably, showed reluctance to duplicate their work, leading to significantly lower or even zero compliance for the steps that lacked hospital interoperability. Consequently, the steps configured without interoperability saw much lower completion rates compared to those that were automatically updated via HL7v2 messages. In this chapter, I will present the two studies separately, one with the compliance of the entire clinical pathway, and the other with a focused analysis of the steps that were configured with interoperability. By doing so, we can better predict the potential compliance rates if the hospital were to implement interoperability across the full clinical pathway. This comparison will offer insight into how streamlining data flows through HL7v2 integration could enhance overall compliance and reduce manual workload for HCPs.

4.2 Metrics

4.2.1 General Compliance

In the compliance assessment of the colorectal cancer clinical pathway studied in the thesis, two distinct sets of data were analysed. Table 4.1 presents the overall results for the entire clinical pathway, while Table 4.2 focuses exclusively on steps configured with

interoperability.

Sessões de pacientes	161
Número total de passos	1305
Ações a menos	582 (45%)
Ações a mais	89
Passos por sessão:	
Min	1
Média	8.1
Max	27
Fitness:	
Min	0.0
Média	0.52
Max	1.0 (24)

Table 4.1: Overall Data for entire clinical pathway.

Sessões de pacientes	161
Número total de passos	896
Ações a menos	192 (21%)
Ações a mais	89
Passos por sessão:	
Min	1
Média	5.6
Max	16
Fitness:	
Min	0.0
Média	0.71
Max	1.0 (50)

Table 4.2: Overall Data of steps with interoperability.

Focusing first on Table 4.1, in total, 161 patient-sessions were analysed, with a cumulative number of 1305 steps recorded across all sessions. On average, each patient session had 8.1 steps, with the number of steps per patient ranging from a minimum of 1 to a maximum of 27. The significant variation in steps per session suggests differing levels of adherence to the pathway, influenced by individual patient conditions and the specific phases of their treatment. Moreover, a deeper analysis revealed that in sessions with fewer steps (1-6) we see the extreme cases, either very high compliance (up to 100%) or low compliance (below 20%). This is easily explained as the smaller processes can either achieve perfect compliance due to easier management, or show poor adherence when key steps are missed. As the number of steps increase (7-14), compliance tends to stabilize in the 40% to 60% range, reflecting partial compliance throughout the sessions, where some steps are regularly skipped. For larger processes (15-27), compliance generally declines,

often falling between 25% and 50%, as it becomes harder to manage and execute a larger, more complex series of steps. This pattern suggests that maintaining high compliance becomes progressively more difficult as the number of steps in a process grows. This explanation was also evident in the second study, as the value of compliance decreased as the number of steps were higher, despite the the number of steps being much lower (896), as well as the range of steps (1-16).

In the 2 tables, it is also visible the number of skipped and inserted activities for both cases. In terms of skipped activities, highlighting portions of the pathway that were not followed, due to treatment adjustments or pathway deviations, in the entire clinical pathway study, were found 582 skipped activities, while on the steps configured with interoperability, there were only 192. This is the first sample, of the many that we will see in the next results, of the difference between these 2 studies as the much higher ratio of skipped activities per total steps (45% for the first, 21% for the second) highlights how much more the skipped activities of the first study affected its fitness results. In terms of inserted events, indicating additional activities done by the HCPs, obviously the 89 events did not change from one study to the other.

Finally, when comparing the fitness values, which measure how closely the actual treatment followed the CPW, the results show a substantial improvement in Table 4.2. The average fitness in Table 1 is 0.52, while in Table 2, it increases to 0.71. This increase in the average fitness score suggests that steps configured solely with interoperability generally exhibit higher compliance, as their completion is automatically recorded, reducing the risk of human error. Besides this, although both tables share the same maximum fitness value of 1.0, with some patient sessions perfectly following the CPW, after a deep analysis, it was revealed that there were only 24 cases of full compliance for the first study compared to the 50 that there were identified in the second, with the number of steps on those sessions ranging from 1 to 7, once again proving that automating the completion of steps through HL7v2 messaging significantly improves adherence to the clinical pathway.

4.2.1.1 Fitness distribution by Compliance levels

Carrying on from the general compliance, here it is shown the distribution of fitness values for each patient session based on compliance levels. The compliance levels are as follows:

- **Low compliance:** [0, 0.2].
- **Below Average compliance:**]0.2, 0.5].
- **Above Average compliance:**]0.5, 0.8].
- **High compliance:**]0.8, 1].

The comparison between Figure 4.1 and Figure 4.2 highlights a clearer shift in the distribution of compliance levels from one study to the other. In study 1, the majority of

patient sessions (82 sessions) fall into the Below Average compliance range, indicating a significant struggle to achieve better adherence to clinical pathways when additional manual updates were required. However, in study 2, this number drops to 40 sessions in this level. Additionally, it shows a substantial increase in the Above Average compliance and High compliance categories. In study 1, 43 sessions reached above-average compliance, while in study 2 this jumps to 66 sessions. Similarly, high compliance results saw an increase from 33 to 53 sessions, reflecting the effectiveness of automated updates in pushing more sessions into higher compliance ranges and emphasizing how the implementation of interoperability led to better adherence to the clinical pathway.

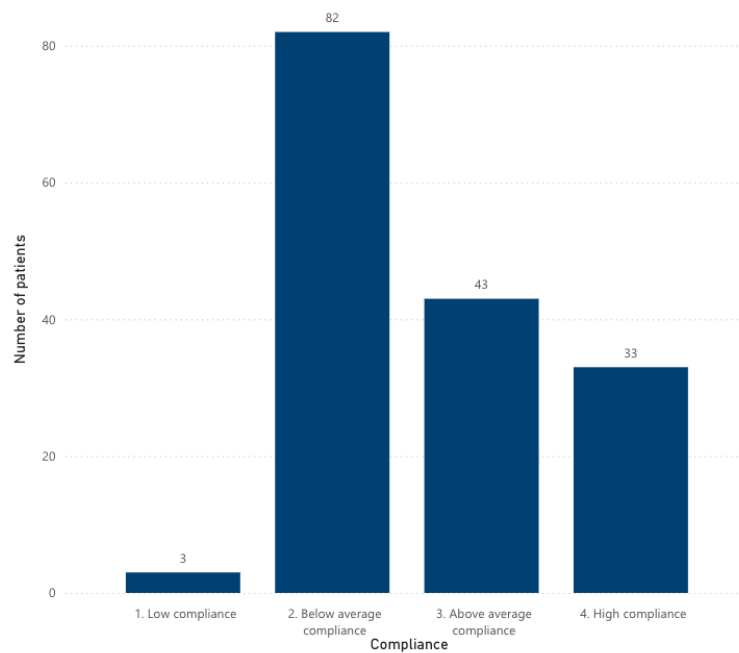


Figure 4.1: Fitness values by compliance levels of entire clinical pathway.

4.2.2 Compliance by Action of the pathway

The table present on Figure 4.3 evaluates the compliance rates for all the actions present in the clinical pathway. The table outlines the total number of steps that were analysed that needed each action to be completed and the consequent number of compliant steps, resulting in a compliance percentage for each action.

The actions are ordered based on their sequence in the pathway. It is important to mention that this, obviously, does not happen for 100% of the cases, as actions can appear in different stages of a patient treatment depending on the severity and the conditions of that specific patient.

This chronological order explains why actions appearing later in the table have lower total counts but higher compliance rates. The last few actions, such as "J) Discharge" and from "M) Oncologic Appointment" to "R) Follow-up appointment" have relatively low total

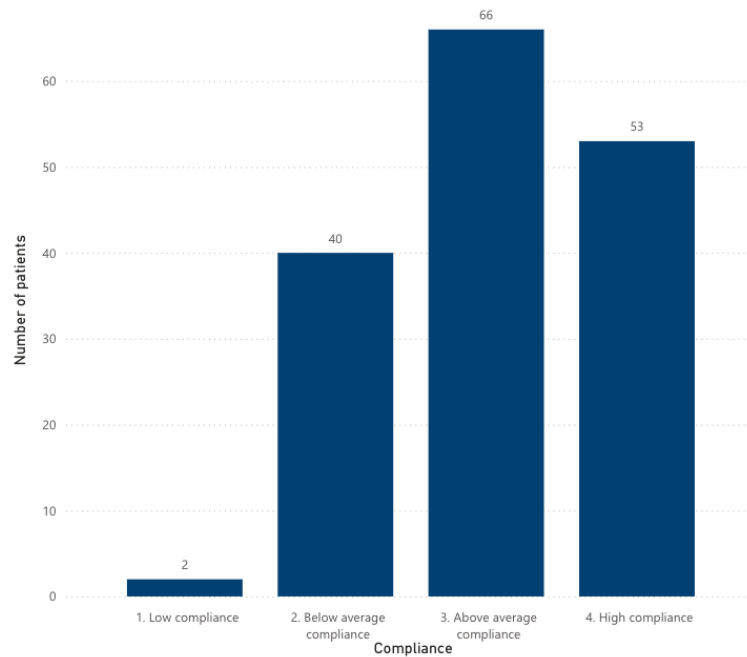


Figure 4.2: Fitness values by compliance levels of steps with interoperability.

counts, indicating that they occur less frequently. However, when these actions do occur, by interoperability, they are always completed, as evidenced by their 100% compliance rates.

In contrast, the actions that occur earlier in the pathway, such as "A) External Appointment" and "C) Schedule Appointment," have higher total counts but lower compliance percentages. For instance, "A) External Appointment" has a compliance rate of 38.27% with a total count of 196, and "C) Schedule Appointment" has a compliance rate of 25.60% with a total of 125 steps. As we saw in the Section 4.2.1, the average number of steps per patient session is 8.1, so these actions occur in nearly all patient sessions, so it becomes hard to maintain a high level of compliance in the process. The lower values can also be explained by the nature of the action in itself. For instance, an "External Appointment" might be scheduled but later canceled or rescheduled, and these updates might not be consistently logged. Additionally, processes like scheduling an appointment or gathering external exam results often involve external systems (e.g., external labs, clinics, etc.), which can lead to delays or partial completions if these external systems are not tightly integrated with the hospital's IT infrastructure.

Another key insight comes from the "Interoperability Configured" column. The lowest compliance percentages in the table align precisely with actions that do not have interoperability configured and are marked as "No". For instance, "B) Complementary Diagnosis Exams" (0.34%) and "G) Pre-op Exams" (0.0%) show the lowest compliance rates and lack interoperability. Additionally, action "H) Operating Room (OR)" has some steps configured with interoperability, and others without it. After an extent analysis to

its steps, it was concluded that all the 34 steps that were not compliant, were also not configured with interoperability. This shows exactly why this is such a factor in these type of mechanisms. When actions rely on additional manual input rather than only the automated system integrations, compliance rates drop significantly, which proves how crucial is the automated data exchange for minimizing the level of errors. This is the reason why it was so important to compare the two studies, as it allowed for a deeper understanding of how interoperability impacts the completion of clinical tasks.

Action	Interop. configured	Compliant count	Total count	Percentage (%)
A) External Appointment	Yes	75	196	38.27
B) Complementary diagnosis exams	No	1	297	0.34
C) Schedule appointment	Yes	32	125	25.60
D) Appointment Scheduling	Yes	68	76	89.47
E) Need of other exams	No	1	7	14.29
F) Pre-op evaluation	Yes	46	53	86.79
G) Pre-op Exams	No	0	45	0.00
H) Operating Room (OR)	*	62	96	64.58
I) Nursing - Stomatherapy	Yes	12	53	22.64
J) Discharge	Yes	52	52	100.00
K) Appointment	Yes	13	49	26.53
L) Subsequent Consultation Colorectal Surgery	Yes	54	54	100.00
M) Oncologic Appointment	Yes	15	15	100.00
N) Consultation	Yes	9	9	100.00
O) Treatment	Yes	8	8	100.00
P) Pathological Anatomy	No	0	9	0.00
Q) OR Admission	Yes	3	3	100.00
R) Follow-up appointment	Yes	3	3	100.00

Figure 4.3: Compliance values by Action of the pathway

4.2.3 Compliance by Phase of the pathway

The stacked area charts in Figures 4.4 and 4.5 presents the compliance percentages for different phases of the colorectal cancer pathway. This is a complement for the previous metric, as all those actions belong to specific phases and both metrics are chronologically ordered. This metric, ultimately, highlights, in a more clear way, the changes in the compliance percentage for each specific stage of a patient treatment.

Looking at the phases, the "Diagnosis" phase in study 1 analysed 619 steps with an average compliance of 38.24%, compared to 409 steps in study 2 with a higher compliance of 52.58%. Similarly, in the "Multidisciplinary team appointment" phase, study 1 examined 157 steps with a compliance of 45.15%, while study 2 reviewed 76 steps and achieved a much higher compliance rate of 88.52%. This increase in percentage are aligned with the takeaways previously established, as in this 2 particular beginning phases there is a mix of actions that have interoperability configured and some that do not. The ones that do not have it make the compliance percentage decrease for the values of the entire clinical pathway assessment. Besides this, the 52.58% value is the lowest for the study 2, with the highest number of steps processed (409) being the obvious reason for it. The "Follow-up

after multidisciplinary team appointment" phase is not really helpfull in terms of insight as it has a very small number of steps analysed, with 6 steps in study 1 and 3 steps in study 2.

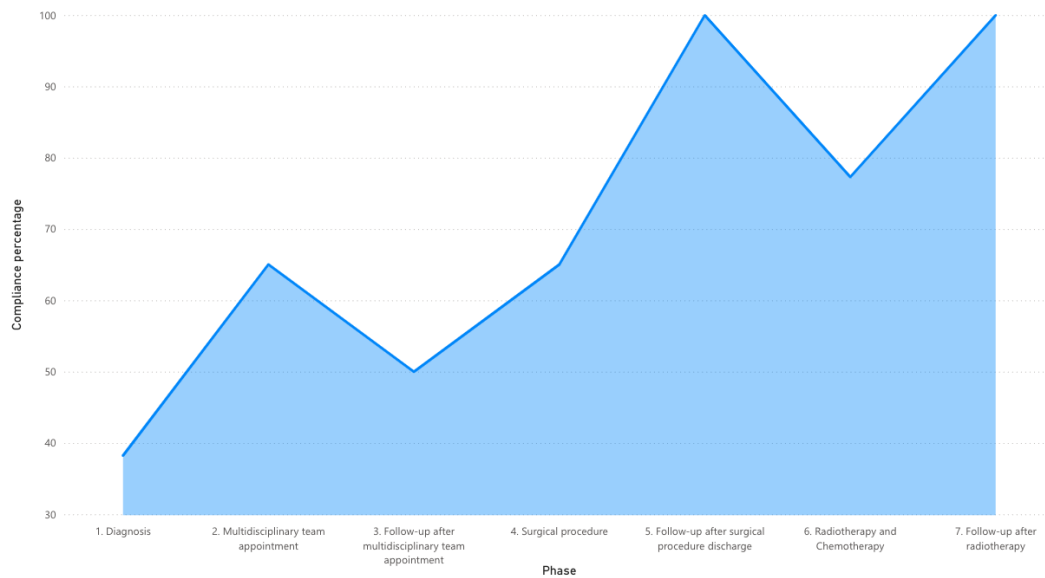


Figure 4.4: Compliance values by Phase of the entire clinical pathway.

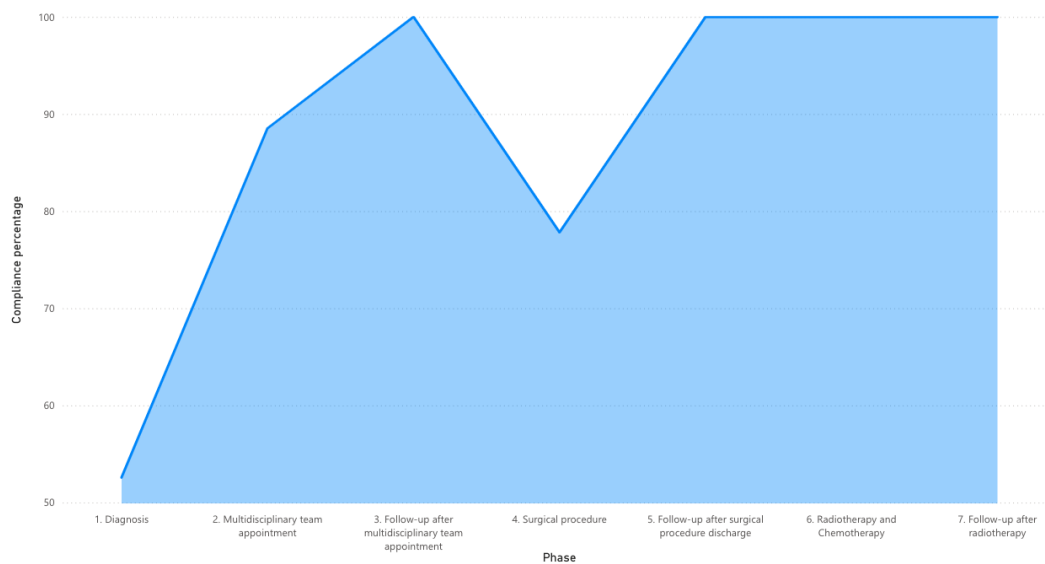


Figure 4.5: Compliance values by Phase of the steps with interoperability.

The "Surgical procedure" phase shows the second worst value of compliance for the study 2, with the compliance going as low as 77.82%. This is directly related, once again, with the increased amount of steps processed in this phase, which is the second highest, after the Diagnosis phase, at 309 steps analysed.

Last, but not least, the last 3 phases of the pathway show the highest values of

compliance for both studies. Study 2 even reaches a perfect compliance of 100% for all three stages. This was also mentioned and explained in the previous metric, as the low amount of steps processed (51, 41 and 9) and importance of these later stages in a patient treatment led to these perfect values. Study 1, however, did not reach 100% of compliance in one of those stages, "Radiotherapy and Chemotherapy", where the compliance value was at 77,29%. This is, once again, the absence of interoperability messages and, consequently, the result of HCPs showing obvious reluctance in duplicate their work, leading to the zero compliance present in the "P) Pathological Anatomy" action present in the Figure 4.3.

4.2.4 Compliance per Week

The line charts in Figure 4.6 and 4.7 provides a visual representation of compliance percentages over the period that the data was being analysed. This timeline starts in mid-July and goes all the way to mid-September 2024.

As we can see from the 2 very similar graphs, the change in compliance over the weeks was quite identical in both studies, with the only difference obviously being the compliance value, where, for the first, it was always around its 52% overall compliance value, while for the second it was around 71%. Apart from this, the behaviour was similar, so the results can be explained together for the two studies.

In the first week we saw values slightly below the average for both, which was due to a significant increase in the company's processing of interoperability messages throughout June and the middle of July and which ended up catching the start of the collection of our compliance data. Then, with the end of this processing, and with a more normalised collection of messages, we saw an increase in the overall number of compliance in the second week compared to the first.

Over the next four weeks, from 25 July to 22 August, the values remained fairly similar, even showing values slightly above the overall compliance value calculated. That week, CHULC began a process of sending out weekly messages in bulk related to different stages of the pathway. This lead to an increase, not only in the number of patient sessions analysed, but in the number of inserted events for the past weeks, as some of those messages are not being used to complete any of the steps of a patient session. Consequently, this influx of unrelated data contributed to the slight decline observed in the compliance chart in recent weeks.

However, despite the slight decrease of 3% and 4%, this is not having a major impact on the values of compliance, but is something to keep an eye on if it continues for much longer.

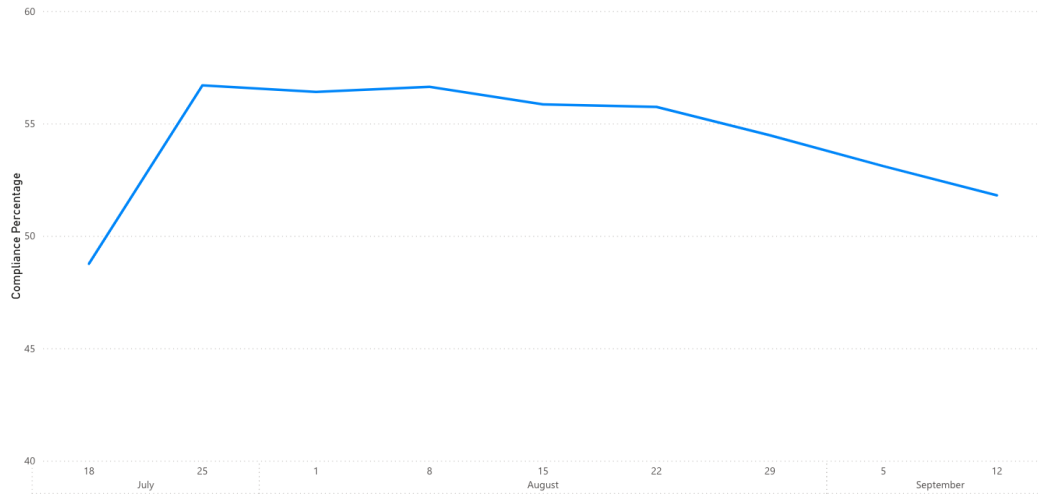


Figure 4.6: Compliance values per Week of the entire clinical pathway.

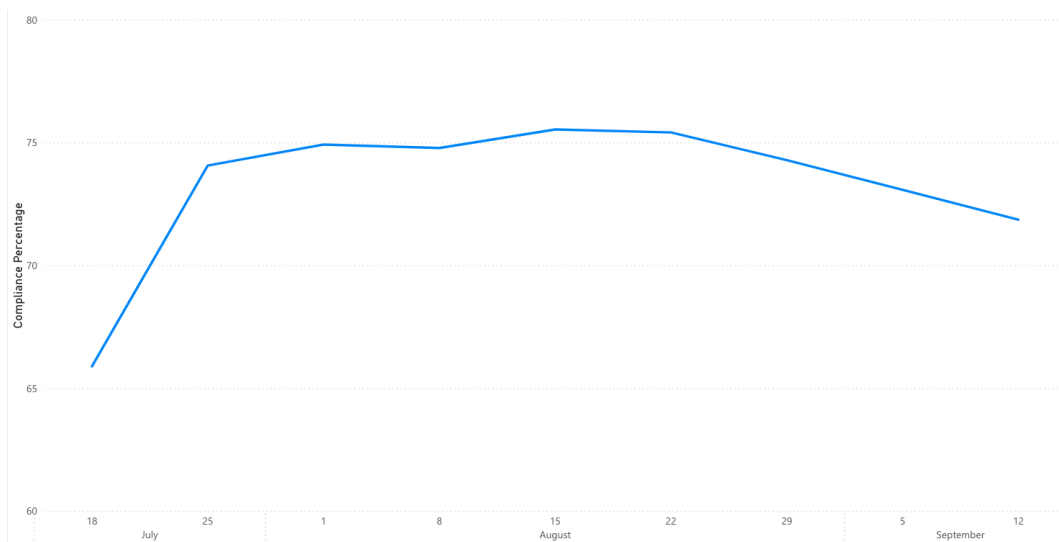


Figure 4.7: Compliance values per Week of the steps with interoperability.

CONCLUSION AND FUTURE WORK

5.1 Conclusion

The primary goal of this thesis was to develop and evaluate a strategy for analysing compliance between CPs and standardized CPWs using EHRs through the implementation of a real-time data streaming solution.

The solution developed in this thesis used several key technological components, each playing a vital role in ensuring its success. DocumentDB, a NoSQL database, was critical in managing the complexity and variability of healthcare data, its flexibility and ability to manage real-time data streams were essential for capturing updates from the Clinical Data Repository (CDR), enabling healthcare organizations to track compliance with clinical pathways using the most current data. MySQL, a relational database, was employed to store structured data and transactional information, such as compliance metrics. Its ability to ensure data consistency and perform complex queries made it ideal for aggregating compliance rates and generating reports. Finally, Power BI was used to transform raw compliance data into actionable insights through interactive visualizations. With its user-friendly dashboards, healthcare administrators could monitor compliance trends, identify areas for improvement, and make informed decisions.

The results generated through the compliance analysis revealed significant variability across different phases of clinical pathways, providing healthcare institutions with a clear picture of how closely their clinical practices adhered to the standardized pathway. This type of insight is very important for healthcare organizations as it allows them to focus on areas that require intervention or resource reallocation.

Moreover, the system's ability to track compliance on a cumulative weekly basis provided institutions with a continuous stream of data that could be used to detect emerging trends, such as improving or deteriorating compliance over time.

Ultimately, the biggest takeaway from the results was that expanding interoperability across the entire clinical pathway could improve the overall compliance and its integration could address the gaps observed in actions where manual input is still required, ultimately leading to more reliable execution of clinical processes.

This solution provides healthcare organizations with a robust and scalable framework for improving compliance with clinical pathways. By offering real-time data capture, flexible data processing and intuitive visualization, the system not only identifies non-compliance but also enables timely interventions that can directly improve patient care and resource efficiency.

It is also important to note that this goes beyond simply tracking compliance to clinical pathways. It acts as a decision support tool that can help healthcare providers optimize their clinical processes, reduce variability in care, and ensure that patients receive the best evidence-based treatments.

5.2 Future Work

In today's rapidly evolving healthcare environment, data-driven insights are becoming essential for improving patient care. So, the effectiveness of every compliance solution improves as more data is collected over time, as extended data collection enables better detection of patterns and trends in clinical pathways. While a two-month period provides valuable insights, analysing data over the course of a year, for example, offers a deeper, more reliable understanding of patient compliance and clinical outcomes. Besides this, some of the work that can be done to improve the system in the future consists of:

1. **Expansion to Other Medical Domains:** While this study focused on colorectal cancer, the methodology could be extended to other clinical domains. Exploring the compliance of CPs across different specialties could help refine the scalability of the framework.
2. **Legal and Ethical Compliance:** With the growing emphasis on data privacy and security, future work could explore integrating compliance with regulations like GDPR into the framework, ensuring that both clinical and legal requirements are met in real time.
3. **Real-Time Feedback for Patients:** Expanding the system to provide patients with real-time updates on their compliance with care plans could further enhance patient engagement, leading to improved outcomes.

BIBLIOGRAPHY

- [1] A. Adriansyah, B. F. van Dongen, and W. M. van der Aalst. “Conformance checking using cost-based fitness analysis”. In: *2011 IEEE 15th International Enterprise Distributed Object Computing Conference*. IEEE. 2011, pp. 55–64 (cit. on pp. 10, 35).
- [2] I. A. Amantea et al. “Semi-automated checking for regulatory compliance in e-Health”. In: *2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)* (2021), pp. 318–325. URL: <https://api.semanticscholar.org/CorpusID:239009557> (cit. on p. 11).
- [3] R. Amor and J. Dimyadi. “The promise of automated compliance checking”. In: *Developments in the built environment* 5 (2021), p. 100039 (cit. on p. 9).
- [4] T. N. Arvanitis. “Semantic interoperability in healthcare.” In: *ICIMTH*. 2014, pp. 5–8 (cit. on p. 23).
- [5] E. Asare, L. Wang, and X. Fang. “Conformance Checking: Workflow of Hospitals and Workflow of Open-Source EMRs”. In: *IEEE Access* 8 (2020), pp. 139546–139566. URL: <https://api.semanticscholar.org/CorpusID:221132094> (cit. on p. 11).
- [6] W. Astriani and R. Trisminingsih. “Extraction, Transformation, and Loading (ETL) module for hotspot spatial data warehouse using geokettle”. In: *Procedia Environmental Sciences* 33 (2016), pp. 626–634 (cit. on p. 14).
- [7] E. Athanasakis. “Medication Safety Practices in Clinical Nursing: Nurses’ Characteristics, Skills, Competencies, Clinical Processes, and Environment.” In: *International Journal of Caring Sciences* 14.3 (2021) (cit. on p. 6).
- [8] D. D. B, S. J. Salim, and S. M. Vargese. “PERFORMANCE EVALUATION OF MYSQL AND MONGODB DATABASES”. In: *International Journal on Cybernetics & Informatics* 5 (2016), pp. 387–394. URL: <https://api.semanticscholar.org/CorpusID:51740255> (cit. on p. 21).
- [9] I. G. Baltopoulos, J. Borgström, and A. D. Gordon. “Maintaining database integrity with refinement types”. In: *European Conference on Object-Oriented Programming*. Springer. 2011, pp. 484–509 (cit. on p. 15).

-
- [10] K. Banker. “MongoDB in Action”. In: 2011. URL: <https://api.semanticscholar.org/CorpusID:60251297> (cit. on p. 19).
- [11] T. Benson and G. Grieve. *Principles of health interoperability: SNOMED CT, HL7 and FHIR*. Springer, 2016 (cit. on p. 24).
- [12] S. Bhartiya and D. Mehrotra. “Challenges and recommendations to healthcare data exchange in an interoperable environment”. In: *electronic Journal of Health Informatics* 8.2 (2014), p. 16 (cit. on p. 23).
- [13] U. Bhat and S. Jadhav. “Moving towards non-relational databases”. In: *International Journal of Computer Applications* 1.13 (2010), pp. 40–47 (cit. on p. 17).
- [14] M. Blount et al. “On the integration of an artifact system and a real-time healthcare analytics system”. In: *Proceedings of the 1st ACM international health informatics symposium*. 2010, pp. 647–655 (cit. on p. 12).
- [15] R. Byali, Jyothi, and M. Shekadar. “Evaluation of NoSQL Database MongoDB with Respect to JSON Format Data Representation”. In: *International Journal of Research Publication and Reviews* (2022-09), pp. 867–871. DOI: [10.55248/gengpi.2022.3.9.24](https://doi.org/10.55248/gengpi.2022.3.9.24) (cit. on p. 17).
- [16] I. Carvalho, F. Sá, and J. Bernardino. “Performance Evaluation of NoSQL Document Databases: Couchbase, CouchDB, and MongoDB”. In: *Algorithms* 16 (2023), p. 78. URL: <https://api.semanticscholar.org/CorpusID:256531386> (cit. on p. 20).
- [17] M. G. Castellanos et al. *Quality-driven ETL design optimization*. US Patent 8,719,769. 2014 (cit. on p. 14).
- [18] J. Celko. “Chapter 11 - Analytic Databases”. In: *Joe Celko’s Complete Guide to NoSQL*. Ed. by J. Celko. Morgan Kaufmann, 2014, pp. 145–171. URL: <https://www.sciencedirect.com/science/article/pii/B978012407192600011X> (cit. on p. 15).
- [19] R. Čerešňák and M. Kvet. “Comparison of query performance in relational a non-relation databases”. In: *Transportation Research Procedia* 40 (2019), pp. 170–177 (cit. on pp. 16, 17, 19).
- [20] A. Chauhan. “A review on various aspects of MongoDB databases”. In: *International Journal of Engineering Research & Technology (IJERT)* 8.05 (2019), pp. 90–92 (cit. on p. 19).
- [21] R. Cruz et al. “Architecture for Intensive Care Data Processing and Visualization in Real-time”. In: *Procedia Computer Science* 184 (2021), pp. 923–928 (cit. on p. 11).
- [22] J. L. Ersek et al. “Clinical pathways and the patient perspective in the pursuit of value-based oncology care”. In: *American Society of Clinical Oncology Educational Book* 37 (2017), pp. 597–606 (cit. on p. 7).

- [23] A. Faraj, B. Rashid, and T. Shareef. "Comparative study of relational and non-relations database performances using Oracle and MongoDB systems". In: *International Journal of Computer Engineering and Technology (IJCET)* 5.11 (2014), pp. 11–22 (cit. on pp. 17, 19).
- [24] B. Feinberg, M. Lingam, and B.-E. Xu. "Mock clinical pathways: A method for exploring the oncology clinical pathway development process". In: *J Clin Pathw* 2 (2016), pp. 1–46 (cit. on p. 7).
- [25] J. Fernandes et al. "Interoperability Between Health Information Systems". In: *International Conference on Sustainability in Software Engineering & Business Information Management*. Springer. 2022, pp. 111–121 (cit. on p. 24).
- [26] M. D. Frutos et al. "Clinical pathway for laparoscopic gastric bypass". In: *Obesity surgery* 17 (2007), pp. 1584–1587 (cit. on p. 2).
- [27] G. Gardarin and M. Melkanoff. "Proving consistency of database transactions". In: *Fifth International Conference on Very Large Data Bases*, 1979. IEEE. 1979, pp. 291–298 (cit. on p. 15).
- [28] V. Gour et al. "Improve performance of extract, transform and load (ETL) in data warehouse". In: *International Journal on Computer Science and Engineering* 2.3 (2010), pp. 786–789 (cit. on p. 14).
- [29] R. Gupta, H. Gupta, and M. K. Mohania. "Cloud Computing and Big Data Analytics: What Is New from Databases Perspective?" In: *Journées Bases de Données Avancées*. 2012. URL: <https://api.semanticscholar.org/CorpusID:27503271> (cit. on p. 15).
- [30] C. Gyorödi, R. Gyorödi, and R. Sotoc. "A comparative study of relational and non-relational database models in a Web-based application". In: *International Journal of Advanced Computer Science and Applications* 6.11 (2015), pp. 78–83 (cit. on pp. 16, 17).
- [31] C. Györödi et al. "A comparative study: MongoDB vs. MySQL". In: *2015 13th international conference on engineering of modern electric systems (EMES)*. IEEE. 2015, pp. 1–6 (cit. on p. 19).
- [32] F. Hak et al. "An exploratory study of a NoSQL database for a clinical data repository". In: *Trends and Innovations in Information Systems and Technologies: Volume 3* 8. Springer. 2020, pp. 476–483 (cit. on p. 22).
- [33] R. Hammami, H. Bellaaj, and A. H. Kacem. "Interoperability of healthcare information systems". In: *The 2014 International Symposium on Networks, Computers and Communications*. IEEE. 2014, pp. 1–5 (cit. on p. 23).
- [34] A. Hamoud, A. S. Hashim, and W. A. Awadh. "Clinical data warehouse: a review". In: *Iraqi Journal for Computers and Informatics* 44.2 (2018) (cit. on p. 13).
- [35] D. H. Havens and L. Boroughs. ""To err is human": a report from the Institute of Medicine". In: *Journal of pediatric health care* 14.2 (2000), pp. 77–80 (cit. on p. 6).

- [36] K. Häyrinen, K. Saranto, and P. Nykänen. “Definition, structure, content, use and impacts of electronic health records: a review of the research literature”. In: *International journal of medical informatics* 77.5 (2008), pp. 291–304 (cit. on p. 5).
- [37] H. Homayouni, S. Ghosh, and I. Ray. “An approach for testing the extract-transform-load process in data warehouse systems”. In: *Proceedings of the 22nd International Database Engineering & Applications Symposium*. 2018, pp. 236–245 (cit. on p. 13).
- [38] Z. Huang et al. “Online treatment compliance checking for clinical pathways”. In: *Journal of medical systems* 38 (2014), pp. 1–14 (cit. on pp. 9, 10).
- [39] R. Hughes. “Patient safety and quality: An evidence-based handbook for nurses”. In: (2008) (cit. on p. 5).
- [40] T. D. Imler, D. J. Vreeman, and J. Kannry. “Healthcare data standards and exchange”. In: *Clinical Informatics Study Guide: Text and Review* (2016), pp. 233–253 (cit. on p. 23).
- [41] N. Jatana et al. “A survey and comparison of relational and non-relational database”. In: *International Journal of Engineering Research & Technology* 1.6 (2012), pp. 1–5 (cit. on pp. 16, 17).
- [42] R. Kumar, S. Charu, and S. Bansal. “Effective way to handling big data problems using NoSQL Database (MongoDB)”. In: *J. Adv. Database Manag. Syst* 2.2 (2015), pp. 42–48 (cit. on p. 19).
- [43] G. C. Lamprinakos et al. “Using FHIR to develop a healthcare mobile application”. In: *2014 4th international conference on wireless mobile communication and healthcare-transforming healthcare through innovations in mobile and wireless technologies (MOBIHEALTH)*. IEEE. 2014, pp. 132–135 (cit. on p. 25).
- [44] N. Leavitt. “Will NoSQL databases live up to their promise?” In: *Computer* 43.2 (2010), pp. 12–14 (cit. on pp. 17, 19).
- [45] E. Liarou et al. “Enhanced stream processing in a DBMS kernel”. In: *International Conference on Extending Database Technology*. 2013. URL: <https://api.semanticscholar.org/CorpusID:11420628> (cit. on p. 15).
- [46] J. M. Lourenço. *The NOVAtthesis L^AT_EX Template User’s Manual*. NOVA University Lisbon. 2021. URL: <https://github.com/joaomlourenco/novathesis/raw/main/template.pdf> (cit. on p. i).
- [47] X. Lu et al. “Conformance checking in healthcare based on partially ordered event data”. In: *Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)*. IEEE. 2014, pp. 1–8 (cit. on p. 9).
- [48] D. Matthews. “What is a clinical pathway?” In: *The Journal of the American College of Dentists* 72.4 (2005), pp. 32–36 (cit. on p. 7).
- [49] M. Mercorella et al. “An architectural model for extracting FHIR resources from CDA documents”. In: *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE. 2016, pp. 597–603 (cit. on p. 25).

- [50] A. Moniruzzaman and S. A. Hossain. “Nosql database: New era of databases for big data analytics-classification, characteristics and comparison”. In: *arXiv preprint arXiv:1307.0191* (2013) (cit. on pp. 19, 20).
- [51] P. M. Nadkarni and C. Brandt. “Data extraction and ad hoc query of an entity—attribute—value database”. In: *Journal of the American Medical Informatics Association* 5.6 (1998), pp. 511–527 (cit. on p. 15).
- [52] M. Novak and K. Rabuzin. “Prototype of a Web ETL tool”. In: *International Journal of Advanced Computer Science and Applications* 5.6 (2014) (cit. on p. 13).
- [53] F. Oemig. “HL7 Version 2. x Goes FHIR.” In: *GMDS*. 2019, pp. 93–98 (cit. on p. 25).
- [54] D. Olteanu and F. Rusu. “Special issue on in-database analytics”. In: *Distributed and Parallel Databases* 35 (2017), pp. 333–334. URL: <https://api.semanticscholar.org/CorpusID:254425009> (cit. on pp. 15, 16).
- [55] D. E. Ong et al. “Usefulness analysis of a Clinical Data Repository design”. In: *2013 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE. 2013, pp. 86–90 (cit. on p. 22).
- [56] T. Patel and T. Eltaieb. “Relational database vs NoSQL”. In: *Journal of Multidisciplinary Engineering Science and Technology (JMEST)* 2.4 (2015), pp. 691–695 (cit. on pp. 16, 17).
- [57] P. A. Pawloski et al. “A systematic review of clinical decision support systems for clinical oncology practice”. In: *Journal of the National Comprehensive Cancer Network* 17.4 (2019), pp. 331–338 (cit. on p. 6).
- [58] J. Pokorný. “Database technologies in the world of big data”. In: *International Conference on Computer Systems and Technologies*. 2015. URL: <https://api.semanticscholar.org/CorpusID:14694321> (cit. on p. 15).
- [59] E. Rojas et al. “Process mining in healthcare: A literature review”. In: *Journal of biomedical informatics* 61 (2016), pp. 224–236 (cit. on p. 6).
- [60] T. Rotter et al. “Clinical pathways as a quality strategy”. In: *Improving healthcare quality in Europe* (2019), p. 309 (cit. on p. 1).
- [61] T. Rotter et al. “Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs”. In: *Cochrane database of systematic reviews* 3 (2010) (cit. on p. 7).
- [62] M. Rovani et al. “Declarative process mining in healthcare”. In: *Expert Systems with Applications* 42.23 (2015), pp. 9236–9251 (cit. on pp. 1, 6, 10).
- [63] A. S. “DATA VISUALIZATION USING POWER BI”. In: *International Scientific Journal of Engineering and Management* (2024). URL: <https://api.semanticscholar.org/CorpusID:269454329> (cit. on pp. 14, 15).
- [64] C. Silva et al. “Business Intelligence for Cardiovascular Disease Assessment”. In: *2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*. IEEE. 2017, pp. 186–193 (cit. on p. 14).

- [65] J. Soliman-Junior et al. "Automated compliance checking in healthcare building design". In: *Automation in construction* 129 (2021), p. 103822 (cit. on p. 9).
- [66] M. Stonebraker, U. Çetintemel, and S. Zdonik. "The 8 requirements of real-time stream processing". In: *ACM Sigmod Record* 34.4 (2005), pp. 42–47 (cit. on p. 12).
- [67] V.-D. Ta, C.-M. Liu, and G. W. Nkabinde. "Big data stream computing in healthcare real-time analytics". In: *2016 IEEE international conference on cloud computing and big data analysis (ICCCBDA)*. IEEE. 2016, pp. 37–42 (cit. on p. 12).
- [68] N. C. Taher et al. "An IoT-Cloud based solution for real-time and batch processing of big data: Application in healthcare". In: *2019 3rd international conference on bio-engineering for smart technologies (BioSMART)*. IEEE. 2019, pp. 1–8 (cit. on p. 12).
- [69] K. Tannir. *RavenDB 2. x beginner's guide*. Packt Publishing Ltd, 2013 (cit. on p. 16).
- [70] N. Thakur and N. Gupta. "Relational and Non Relational Databases: A Review". In: *Journal of University of Shanghai for Science and Technology* 23.8 (2021), pp. 117–121 (cit. on pp. 16, 17).
- [71] A. Torab-Miandoab et al. "Interoperability of heterogeneous health information systems: a systematic literature review". In: *BMC Medical Informatics and Decision Making* 23.1 (2023), p. 18 (cit. on pp. 24–26).
- [72] J. Van de Klundert, P. Gorissen, and S. Zeemering. "Measuring clinical pathway adherence". In: *Journal of biomedical informatics* 43.6 (2010), pp. 861–872 (cit. on p. 2).
- [73] A. Villar et al. "Integrating and analyzing medical and environmental data using ETL and Business Intelligence tools". In: *International journal of biometeorology* 62 (2018), pp. 1085–1095 (cit. on p. 13).
- [74] J. Wahyudi et al. "Database Management Education in MYSQL". In: *Edumaspul: Jurnal Pendidikan* (2022). URL: <https://api.semanticscholar.org/CorpusID:258806202> (cit. on pp. 20, 21).
- [75] A. Walinjkar and J. Woods. "FHIR tools for healthcare interoperability". In: *Biomedical Journal of Scientific and Technical Research* 9.5 (2018) (cit. on p. 25).
- [76] H. J. Watson. "Tutorial: Big data analytics: Concepts, technologies, and applications". In: *Communications of the Association for Information Systems* 34.1 (2014), p. 65 (cit. on p. 15).
- [77] M. Weidlich et al. "Process compliance analysis based on behavioural profiles". In: *Information Systems* 36.7 (2011), pp. 1009–1025 (cit. on p. 2).
- [78] L. Welling and L. Thomson. "MySQL Tutorial". In: 2003. URL: <https://api.semanticscholar.org/CorpusID:67268356> (cit. on p. 21).
- [79] J. Zakir, T. Seymour, and K. Berg. "Big data analytics." In: *Issues in Information Systems* 16.2 (2015) (cit. on p. 17).

BIBLIOGRAPHY

- [80] Z. Zhang et al. "Capabilities of rule representations for automated compliance checking in healthcare buildings". In: *Automation in Construction* 146 (2023), p. 104688 (cit. on p. 9).



2024

Analysis of electronic health records to assess compliance with Clinical Pathways

David Delgado