UNISINOS

Graduate Program in

Applied Computing

Academic Doctorate

Blanda Helena de Mello

A semantic interoperability model based on NLP for non-structured health data

São Leopoldo, 2025

Blanda Helena de Mello

A SEMANTIC INTEROPERABILITY MODEL BASED ON NLP FOR NON-STRUCTURED HEALTH DATA

Doctoral Thesis presented as a partial requirement to obtain the Doctoral's degree from the Applied Computing Graduate Program of the Universidade do Vale do Rio dos Sinos — UNISINOS

Advisor:
Prof. Sandro José Rigo, PhD

Co-advisor:
Prof. Cristiano André da Costa, PhD

São Leopoldo
2025

M527s Mello, Blanda Helena de.
A semantic interoperability model based on NLP for non-structured health data / Blanda Helena de Mello. – 2025.
168 f. : il. ; 30 cm.

Tese (doutorado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, 2025.
"Orientador: Prof. Dr. Sandro José Rigo
Coorientador: Prof. Dr. Cristiano André da Costa"

1. Electronic health record. 2. Machine learning. 3. Natural language processing. 4. Ontology. 5. Semantic interoperability. I. Título.

CDU 004.4

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecária: Silvana Dornelles Studzinski – CRB 10/2524)

(This page was left blank intentionally. It is only used to replace the real approval page.)

To my family.

"All we have to decide is what to do with the time that is given us."
— The Fellowship of the ring — J.R.R.Tolkien

ACKNOWLEDGEMENTS

First and foremost, I want to express my deepest gratitude to my advisor, Professor Sandro José Rigo. Your guidance and insightful suggestions made all the difference in this work. Having you as a mentor and professor has been an honor. I am also profoundly grateful to Professor Cristiano André da Costa for accepting the role of co-advisor on this journey. Your time, wisdom, and leadership in the SOFTWARE LAB research group were fundamental to my academic growth.

A special thank you to Professor Bjoern Eskofier for welcoming me into your lab in Germany for six months. This experience gave me a fresh perspective on research and professional relationships. To my MaD Lab and LME Lab colleagues, your friendship turned my time abroad into something that felt like home. To my dear lab mates — Marlies Nitschke, Maike Stöve, Hamid Moradi, Ann-Kristin, Rebecca Lennartz, Anne Koelewijn, Fatemeh Salehi, Charly, Chuyi Wang, Sophie Fleischmann, Katharina Jaeger, Richard Dirauf, René Raab, Kai Kleide, and Simon Dietz — you made the everyday moments in the lab so special. And to my LME Lab mates who received me with such warmth — Paula Perez, Frauke Wilm, Maximilian Rohleder, Tomás Arias-Vergara, Noah Maul, Aniol, Mareike Thies, Annette Schwarz, Nora Gourmelon, Prathmesh Madhu, and the kindest friend, Nastassia Vysotskaya — thank you for your kindness.

To the professors of the Graduate Program in Applied Computing (PPGCA) and the MyDigitalHealth project, thank you for your passion and dedication. You inspired me to aim higher and dream bigger. A special thank you to Professor Marta Bez — your unwavering support, wise advice, and care as both a professor and a friend have meant the world to me.

To my research group mates, thank you for being there through all the ups and downs, especially during the challenges of the pandemic. To Felipe Zeiser, my PhD mate and dear friend, thank you for sharing this crazy journey with me — the struggles, the laughs, the adventures abroad, and the beers in Germany. Here's to many more adventures. To Diego Pinheiro and Luana Rockenback, your love and friendship have been a constant anchor in my life. Bruna Donida and Ana Alegretti, meeting you both was such a gift! Our shared laughter and moments made everything better. Lara Liz Freire, your friendship and encouragement kept me going even on my most challenging days. Flora Morais, my dear friend and now a resident of my heart, thank you for being you and saving my images for this thesis! Adrielly, thank you for being by my side, supporting me, and cheering me on during the last chaotic month. Isadora Dias, last but certainly not least, welcome to the ship of crazy PhD candidates; I'll be waiting for you on the other side.

To the Google Women Techmakers project and all the volunteers, my time as an ambassador brought me experiences and lessons I will carry with me forever. I am also grateful to the Ministry of Health, the Digital Health Secretary (SEIDIGI), and DATASUS for giving me the opportunities and insights that shaped this research, especially Paula Xavier and Robson Matos.

To my family, your unconditional love and support have been my bedrock throughout this journey, fueling my perseverance and success. All my love to my mother, father, grandma, and siblings. And above all, I thank God for life's grace and blessings.

The authors would like to thank the Coordinating Agency for Advanced Training of Graduate Personnel (CAPES) (C.F. 001), and National Council for Scientific and Technological Development (CNPq) (No. 309537/2020-7).

ABSTRACT

The healthcare domain faces significant challenges in managing the rapidly growing volume of data generated daily, particularly in the collection and sharing of this information. Healthcare professionals such as physicians, nurses, radiologists, cardiologists, surgeons, and other specialists frequently enter patient data into electronic systems, often in an open, unstructured textual format. We conducted a literature review that reveals several challenges in processing real-world data, with one critical issue being the scarcity of tools and dictionaries available in Portuguese for the healthcare sector. This gap, coupled with the unique challenges inherent in healthcare data processing, adds considerable complexity to extracting and structuring essential information from clinical records. Additionally, ensuring data interoperability between different healthcare providers becomes challenging when these providers do not initially aim for interoperability during input data. Observing these challenges, this research proposed a model to enable semantic interoperability of clinical notes from electronic health record systems. The methodology used in this research has an applied and exploratory character, and it has been evaluated through the development of a prototype. This approach aims to address some of the current limitations in data processing and integration, specifically within the Portuguese healthcare context, and to create a flexible model that can treat real-world data more effectively in structuring and sharing data. This research is part of the MyDigitalHealth project, a collaboration between the university and six hospitals in Porto Alegre, which provided data from hospitalized patients who tested positive for COVID-19, ensuring a real-world context for data issues. We analyzed the characteristics of the data with respect to interoperability between providers and proposed a model that involves hybrid techniques for information extraction, lexical normalization, and structure for standard harmonization. Thus, we defined a set of experiments using machine learning, combining the Transformers architecture for entity recognition with natural language processing for lexical normalization and semantic matching and adopting OWL ontologies as an intermediary representation structure. The experiments revealed three main contributions. First, we developed a specialized annotated dataset, classifying six entities with 18,666 validated annotations by specialists in 314 documents. Second, we conducted experiments using BERT models fine-tuned on our small dataset for entity recognition, achieving 95% accuracy, with precision rates of 90% for classifying entities related to Invasive or Therapeutic Procedures and 89% for Disease or Syndrome and Diagnostic Procedures. These results demonstrate the model's effectiveness in extracting relevant information from unstructured clinical notes. Third, ontologies as intermediary representation structures ensured semantic consistency and enhanced interoperability in an independent format. The limitations and opportunities for future studies from this research include applying the model to data from different domains, such as nursing notes, odontology, clinic context, and accountability records. Another topic is the gap in term disambiguation and semantic alignment in healthcare data, focusing on linking terminologies to structured data, ensuring international coding for clinical data, and enabling interoperability across borders. Finally, this research aims to contribute to the continuity of citizen healthcare and guide developers and providers in building robust and complex platforms that implement the use of healthcare standards. We also expect more and more professionals and health managers to improve healthcare worldwide through the adoption of international standards within electronic health record systems.

Keywords: Semantic interoperability. Electronic health record. Ontology. Natural Language Processing. Machine Learning.

RESUMO

O domínio da saúde enfrenta desafios significativos no gerenciamento do crescente volume de dados gerados diariamente, particularmente na coleta e compartilhamento dessas informações. Profissionais de saúde, como médicos, enfermeiros, radiologistas, cardiologistas, cirurgiões e outros especialistas frequentemente inserem dados de pacientes em sistemas eletrônicos, geralmente em um formato textual aberto e não estruturado. Uma revisão da literatura revela vários desafios no processamento de dados do mundo real, com uma questão crítica sendo a escassez de ferramentas e dicionários disponíveis em português para o setor de saúde. Essa lacuna, juntamente com os desafios únicos inerentes ao processamento de dados de saúde, adiciona considerável complexidade à extração e estruturação de informações essenciais de registros clínicos. Além disso, garantir a interoperabilidade de dados entre diferentes provedores de saúde se torna desafiador quando esses provedores não visam inicialmente a interoperabilidade durante a coleta de dados. Observando esses desafios, esta pesquisa propôs um modelo para permitir a interoperabilidade semântica de notas clínicas de sistemas de prontuários eletrônicos. A metodologia usada nesta pesquisa tem um caráter aplicado e exploratório, e foi avaliada por meio do desenvolvimento de um protótipo. Esta abordagem visa mapear as limitações atuais no processamento e integração de dados, especificamente no contexto de notas clínicas em português brasileiro, e criar um modelo flexível que possa tratar dados do mundo real de forma mais eficaz na estruturação e compartilhamento de dados. Esta pesquisa faz parte do projeto MinhaSaudeDigital (MSD), uma colaboração entre a universidade e seis hospitais de Porto Alegre, que forneceram dados de pacientes hospitalizados que testaram positivo para COVID-19, garantindo um problema do mundo real para o contexto de dados de saúde. Foram analisadas as características dos dados com relação à interoperabilidade entre provedores e proposto um modelo que envolve técnicas híbridas para extração de informações, normalização lexical e estruturação de dados para harmonização de padrões. Assim, definiu-se um conjunto de experimentos que emprega o aprendizado de máquina, combinando a arquitetura Transformers para reconhecimento de entidades com processamento de linguagem natural para normalização lexical e correspondência semântica, por adotando ontologias OWL como uma estrutura de representação intermediária. Os experimentos revelaram três contribuições principais. Primeiro, o desenvolvimento de um conjunto de dados anotados especializados, classificando seis entidades com 18.666 anotações validadas por especialistas em 314 documentos. Em segundo lugar, conduzidos experimentos usando modelos BERT ajustados em um pequeno conjunto de dados para reconhecimento de entidades, alcançando 95% de precisão, com taxas de precisão de 90% para classificar entidades relacionadas a Procedimentos Invasivos ou Terapêuticos e 89% para Doenças ou Síndromes e Procedimentos Diagnósticos. Esses resultados demonstram a eficácia do modelo na extração de informações relevantes de notas clínicas não estruturadas. Terceiro, ontologias como estruturas de representação intermediárias garantiram a consistência semântica necessária à interoperabilidade mantendo um formato independente. As limitações e oportunidades para estudos futuros desta pesquisa incluem a aplicação do modelo a dados de diferentes domínios, como notas de enfermagem, odontologia, contexto clínico e registros de responsabilidade. Outro tópico é a lacuna na desambiguação de termos e alinhamento semântico em dados de saúde, com foco na vinculação de terminologias a dados estruturados, garantindo codificação internacional para dados clínicos e permitindo a interoperabilidade entre fronteiras.

Finalmente, esta pesquisa visa contribuir para a continuidade do cuidado e saúde do cidadão e orientar desenvolvedores e provedores na construção de plataformas robustas e complexas que implementem o uso de padrões de saúde. Também espera-se que cada vez mais profissionais e gestores de saúde melhorem a assistência médica em todo o mundo por meio da adoção de padrões internacionais em sistemas de prontuários eletrônicos.

Palavras-chave: Interoperabilidade Semântica. Registro Eletrônico de Saúde. Ontologia. Processamento de Linguagem Natural. Aprendizado de Máquina.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ANSI	American National Standards Institute
BERT	Bidirectional Encoder Representations from Transformers
BERT-GT	Bidirectional Encoder Representations from Transformers - Graph Transformer
Bi-GRU	Bidirectional-Gated Recurrent Unit
Bi-LSTM	Bidirectional-Long Short-Term Memory
Bi-LSTM-CRF	Bidirectional-Long Short-Term Memory-Conditional Random Field
BioBERT	Biomedical Bidirectional Encoder Representations from Transformers
BioNER	Biomedical Named Entity Recognition
BioRE	Biomedical Relation Extraction
BioTop	BioTopLite2
CDE	Common Data Element
CDSS	Clinical Decision Support System
ChEBI	Chemical Entities of Biological Interest
CIM	Clinical Information Model
CKM	Clinical Knowledge Manager
CNN	Convolutional Neural Network
COVID-19	Coronavirus disease 2019
CRF	Conditional Random Field
CSV	Comma-separated values
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
EHR	Electronic Health Record
EL	Entity Linking
Elmo	Embeddings from Language Models
EN	Entity Normalization
FHIR	Fast Healthcare Interoperability Resources
FMA	Foundational Model of Anatomy Ontology
GraphSAGE	Graph SAmple aggreGatE
GRU	Gated Recurrent Unit
GTP	Generative Pre-Training
HIMSS	Healthcare Information and Management Systems Society
HIPAA	Health Insurance Portability and Accountability Act

HITECH	Health Information Technology for Economic and Clinical Health Act
HL7	Health Level Seven International Foundation
HL7 CCD	Continuity of Care Documento
HL7 CCR	Continuity of Care Record
HL7 CDA	Clinical Document Architecture
HL7 RIM	Reference Information Model
HTML	Hyper Text Markup Language
ICD-11	International Classification of Diseases-version 11
ICD-10	International Classification of Diseases-version 10
ICD-9	International Classification of Diseases-version 9
IE	Information Extraction
IHE	Integrating the Healthcare Enterprise
IOT	Internet of Things
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
LGPD	General Data Protection Law
LOD	Linked Open Data
LOINC	Logical Observation Identifiers Names and Codes
LSTM	Long Short-Term Memory
MAGNN	Metapath Aggregated Graph Neural Network
M-BERT	BERT-Base-Multilingual
MedDRA	Medical Dictionary for Regulatory Activities
MeSH	Medical Subject Headings
ML	Machine Learning
MLM	Masked Language Model
MLP	Multi Layer Perceptron
MS	Ministry of Health
MS/GM	Ministry of Health/Minister's Office
MSD	Minha Saude Digital
N3	Notation 3
NED	Named Entity Disambiguation
NER	Named Entity Recognition
NLP	Natural Language Processing
NSP	Next Sentence Prediction

OWL	Ontology Web Language
OWL-DL	Ontology Web Language-Description Logid
OWLS	Ontology Web Language System
RDF	Resource Description Framework
RE	Relation Extraction
REST	Representational State Transfer
R-GCNs	Relational Graph Convolutional Networks
RM	Reference Model
RNN	Recurrent Neural Network
RNDS	National Healthcare Data Network in Brazil
RTF	Rich Text Format
SKOS	Simple Knowledge Organization System
SNOMED-CT	Systematized Nomenclature of Medicine-Clinical Terms
SPARQL	SPARQL Protocol and RDF Query Language
SVM	Support Vector Machine
SWRL	Semantic Web Rule Language
TSV	Tab-Separated Values
TL	Transfer Learning
W3C	World Wide Web Consortium
WHO-ATC	World Health Organization-Anatomical Therapeutic Chemical
WSD	Word Sentence Disambiguation
XML	Extensible Markup Language

CONTENTS

1 INTRODUCTION

"The right answer doesn't matter at all: the essential thing is asking the right questions."
Mario Quintana

Digital transformation in healthcare has become a key process to ensure the sustainability of the care environment. According to Woods et al. (2023), digital health refers to using technology to provide and support health care services. In that way, and considering the current scenario, the aging population and health inequities are growing, and digital health can provide the opportunity to improve care at scale, reducing sizeable costs. As highlighted by Mougin, Hollis and Soualmia (2022), digital health can be pivotal in supporting reducing health inequalities and disparities.

The COVID-19 pandemic has introduced several challenges for the scientific community and healthcare services worldwide. These challenges include recording, managing, and monitoring patients' clinical data at different healthcare levels and fostering effective, integrated, and secure procedures (SIM et al., 2023; HEYMANN; SHINDO, 2020). An efficiently interoperable healthcare system is crucial in combating these threats (GANSEL; MARY; BELKUM, 2019) and led the development of methods to collect, standardize, disseminate, and reuse health data globally (MOUGIN; HOLLIS; SOUALMIA, 2022).

The growing need to integrate systems, exchange data, and collect high-quality health information (SIM et al., 2023; EVANS, 2016) represents a pathway to reducing unnecessary care, preventing harm, improving patient engagement through a patient-centered approach, and enhancing opportunities for monitoring, risk management, and quality improvement (WOODS et al., 2023). This has become a reality as more and more healthcare establishments have integrated the use of Electronic Health Records (EHR) Mougin, Hollis and Soualmia (2022) into the routine of healthcare professionals.

An Electronic Health Record (EHR) is a system designed to record healthcare data in a processable manner, enabling secure storage and communication, facilitating data sharing, and preserving patient information efficiently and with high quality throughout a patient's life ISO18308 (2011). The purpose of an EHR is to emphasize the need for quality and standard adoption without mandating or specifying which standard to adopt. The appropriate standard should be defined based on the specific needs for data sharing and storage (ROEHRS et al., 2018). Additionally, ensuring data sharing in the healthcare domain involves addressing well-defined requirements related to data issues, data ownership, growing support for data sharing, data sharing initiatives, and how

federated data might be applied as a solution (HULSEN, 2020).

The transition from paper-based to electronic systems presents various challenges, with the need to field-code data at the point of care for electronic handling being a major barrier to acceptance by physicians and nurses, as they lose the ease of natural language entry and understanding that handwritten notes or even electronic textual notes provide (RAGHUPATHI; RAGHUPATHI, 2014; MARTIN-SANCHEZ; VERSPOOR, 2014).

In addition to implementing Electronic Health Records (EHR), as mentioned by Shull (2019), organizations need to consider coding and semantics (centralized and standardized vocabularies), privacy issues, and associated costs. As a result, there has been significant attention towards proposing collaborative measures for digital transformation in the healthcare sector (WOODS et al., 2023).

According to Lazarova et al. (2022), the healthcare landscape is shifting towards preventive care and early diagnosis. Healthcare data are inherently fragmented, as they are collected across different levels of care, services, and specialties and are even more heterogeneous due to the involvement of both public and private sectors. In this heterogeneous scenario, enabling the exchange of these diverse data among healthcare establishments and the government represents a key challenge, often related to interoperability. (SIVARETHINAMOHAN; SUJATHA; BISWAS, 2021; GAGALOVA et al., 2020; XIAO; CHOI; SUN, 2018).

Despite the positive adoption of EHR systems in the healthcare domain, the full potential of these systems is often hindered by challenges such as interoperability issues, user resistance, high implementation costs, and the need for ongoing training and support. The lack of interoperability is a significant obstacle to data sharing between institutions, regions, or countries (LEE; KIM; LEE, 2021; NEGRO-CALDUCH et al., 2021). It is critical in advancing toward a fully integrated digital health system.

As highlighted by Benson and Grieve (2021), when two or more different systems exchange information, regardless of the technologies involved, we call it interoperability. An electronic health record aims to share health information between different healthcare providers and to guarantee interoperability by the semantic level of health data, and it adopts a set of standards.

To meet these needs, semantic interoperability in the healthcare domain incorporates standards to ensure high-quality data exchange, preserving both context and meaning (ALEXOPOULOS, 2020), by adopting terminologies, vocabularies, and codable terms to establish consensus on the meaning of data (PALOJOKI et al., 2024).

However, healthcare providers often develop EHR solutions that adopt proprietary standards, making data integration and interoperability between different organizations complex (ROEHRS et al., 2018). In this scenario, sharing high-quality data using healthcare standards enables the primary use of information for ongoing care. It supports secondary purposes such as research, quality improvement, and policy-making.

As reinforced by D'Amore et al. (2021), Gansel, Mary and Belkum (2019), the need for qualified data for secondary usage is crucial to support medical and clinical research, epidemiological surveillance, and patient care management (across hospitals, general practitioners, nursing, and physiotherapy).

The Brazilian Unified Health System, called "Sistema Único de Saúde" (SUS), was created in 1988. Since then, Brazil has been the only country with over 200 million citizens, with a universal health system used by Brazilians, which is exclusively used by 75% of the population, free of charge (PAIM et al., 2011), while the last 25% also use supplementary health sector. However, the SUS healthcare units lack integrated Electronic Health Records nationally. Each unit has the autonomy to manage their patients' history, choosing which systems they wish to use and standards.

We highlighted the publication of the General Data Protection Law[1] an important effort forward to data sharing and interoperability, which aims to address the definitions of different types of data, levels of sensitivity, and how they should be handled, ensuring respect for privacy, inviolability, and personal rights. The General Data Protection Law (LGPD) guides how new studies using real-world data interact with this information.

In this context, recognizing the inevitability and necessity of adopting digital systems in the health domain, Brazil has proposed an integrated network for receiving, consolidating, and consulting data. In 2020, the Ministry of Health (MoH), through Ordinance No. 1,434 of May 28, 2020, Brazil instituted the National Healthcare Data Network (RNDS), the national platform to consolidate and reunite data from different healthcare public and private health establishments. Through Ordinance GM/MS Nº. 1,046 of May 24, 2021, during the pandemic, it was established the norm that mandates the reporting of COVID-19 Clinical Laboratory Diagnosis - Rapid Test Profiles results from public, private, university, and other laboratories across the national territory to the RNDS.

The project RNDS became a prominent initiative to respond to pandemics and a strategic digital health integrating systems to interoperate data. Nonetheless, MoH has been promoting efforts to improve health informatics and digital health since 2011 by the Ordinance Nº. 2,073, of August 31, regulates the use of health information and interoperability standards for health information systems within the Unified Health System (SUS) at the Municipal, District, State, and Federal levels, as well as recommendations for private systems and the supplementary health sector.

Also, to improve data qualification and empower patients as the owners of their information, Brazil established a unified identification for citizens using the CPF code, the Individual Taxpayer Registry, in Portuguese 'Cadastro de Pessoas Físicas' (CPF), through Law No. 14,534 enacted on January 11, 2023. This facilitates EHR systems as

[1]General Data Protection Law (Lei Geral de Proteção de Dados Pessoais, in Portuguese) <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>

a universal identifier for patients and allows RNDS to define a Master Patient Index (MPI), enhancing data interoperability at the federal level.

The RNDS project continues to address the initiative of promoting technical resources to meet the emerging need for interoperability, advocating for and adopting HL7 Fast Healthcare Interoperability Resources (FHIR) as the official standard for data exchange in the RNDS. Legally, the public health sector must follow the directives from the Ministry of Health (MoH); however, the private sector, unless mandated, may consider these recommendations and adapt to the information models defined in the national standard. Therefore, this initiative represents a significant first step towards adopting a national standard and creating an integrated system for consolidating and consulting health data.

The electronic medication prescribing definition was recently published ensuriing the transformation objective, defined through the Ordinance SAES/MS Nº 50, of February 9, 2022. The norm establishes the Clinical Information Model (CIM) of Medication Prescription Record and Medication Dispense Record. A significant achievement, whereas the effects on the medication error rate showed a significant relative risk reduction of 13% to 99% for medication errors through an electronic prescribing medication (computerized physician order entry systems) (AMMENWERTH et al., 2008).

In recent years, Brazil has taken on a national regulatory role by defining policies focused on the transformation toward digital health. According to Decree No.11,358[2] of January 1st, 2023, Brazil created the Secretariat of Information and Digital Health (in Portuguese, Secretaria de Informação e Saúde Digital - SEIDIGI), responsible for overseeing technological and semantic decisions to develop, integrate, and ensure the interoperability of information and communication in digital health and telehealth within the Unified Health System (SUS) environment.

Desiring to overcome the challenges to achieve an EHR interoperable, according to Kah and Zeroual (2021), Malm-Nicolaisen et al. (2019), Adel et al. (2019), interoperability consists of combining controlled vocabulary codes, Clinical Information Model (CIM), terminologies, and health standards adoption. Over the years, experts in the medical domain manually extracted patient data to meet this goal of interoperability (KOLECK et al., 2019). With the EHR adoption, this scenario provokes a growing volume of healthcare data, which has been the hallmark of this digital transformation (NASEEM et al., 2020). However, with this large volume of data, manual activities become impractical (MARTIN-SANCHEZ; VERSPOOR, 2014; KOLECK et al., 2019).

Using an EHR presents challenges as much of the data is in proprietary formats, lacks formal standards, and is unstructured or inadequately structured (resulting in

[2]DECRETO Nº 11.358, de 01 de janeiro de 2023)<https://www.planalto.gov.br/cciviL_03////_Ato2023-2026/2023/Decreto/D11358.htm>

complex, heterogeneous databases that hinder novel research) (ROEHRS et al., 2017). Heterogeneous repositories make it difficult to identify and link data, as they come from different interactions (e.g., imaging exams, medical requests, prescriptions, reports, medications, and respective dosages) (SIVARETHINAMOHAN; SUJATHA; BISWAS, 2021; XIAO; CHOI; SUN, 2018). Furthermore, some studies have proposed integrating heterogeneous sources and adopting health standards to achieve interoperability (SÁEZ et al., 2013; MALDONADO et al., 2020; LEGAZ-GARCÍA et al., 2016; LEZCANO; SICILIA; RODRÍGUEZ-SOLANO, 2011), and have also explored integrated data for other research (GAGALOVA et al., 2020), including secondary uses such as cohort identification, public health management, pharmacovigilance, Clinical Decision Support Systems (CDSS), medication management, and clinical text summarization (HOUSSEIN; MOHAMED; ALI, 2021).

Advances in the health domain, specifically related to the EHR systems adoption, have converged to the emergence of new areas related to health, such as blockchain technology (ROEHRS et al., 2018), Clinical Information Models and healthcare standards (PALOJOKI et al., 2024; LEE; KIM; LEE, 2021; ADEL et al., 2019; ROEHRS et al., 2019), Information Extraction (IE), Natural Language Processing (NLP), and Machine Learning (ML) methods (NEGRO-CALDUCH et al., 2021; XIAO; CHOI; SUN, 2018). This scenario opened new health domain research paths, which promotes interdisciplinary research challenges concentrated on interfaces between human languages, natural language processing (NLP), and machine learning techniques (SHEN; HSIA; HSU, 2021). In order to perform information extraction activities, the state-of-the-art points to some approaches employing Natural Language Processing (NLP) and Machine Learning (ML) techniques (XIAO; CHOI; SUN, 2018). According to the study conducted by Houssein, Mohamed and Ali (2021), Li et al. (2021) the most used approaches to information extraction were Named Entity Recognition (NER) and Relation Extraction (RE).

On the one hand, NER and RE activities showed promising results when using ML approaches with RNN (HOUSSEIN; MOHAMED; ALI, 2021) architectures such as LSTM, BiLSTM, and hybrids Bi-LSTM-CRF (GIORGI; BADER, 2020; RIVERA-ZAVALA; MARTÍNEZ, 2021), GRU (ZHANG et al., 2018), and BI-GRU(HONG et al., 2020) and CNN. In contrast, although newly released, the transformers architectures (DEVLIN et al., 2018) have shown excellent results for the information extraction application using BERT models (Bidirectional Encoder Representations from Transformers) on downstream tasks NER. The model excels in applying NER and RE tasks, which allows exploring the same pre-trained model and fine-tuning the desired task.

The studies that showed promising results in IE proposed hybrid approaches combining ML and NLP. However, there is a need to advance information extraction

tools that consider the semantic context in which they operate. Furthermore, qualified data enables reusability and facilitates information sharing, significantly improving patient care. According to Salomi, Claro et al. (2020), a qualified EHR system helps eliminate rework, reduce errors, and promote individualized patient care from a patient-centered perspective. From a citizen's perspective, it supports the creation of primary care initiatives, disease control and monitoring, cost reduction, and increased effectiveness through citizen engagement. Additionally, it enables the delivery of high-quality care and the allocation of essential resources (KIM; JOSHI, 2021). As reinforced by Adel et al. (2019), Lee, Kim and Lee (2021), achieving semantic interoperability involves adopting controlled standards and vocabularies to maintain the same meaning across all providers involved in generating or receiving data, regardless the technological choices.

This research originates from the MyDigitalHealth project, approved by the Coordination for the Improvement of Higher Education Personnel (CAPES) and the National Council for Scientific and Technological Development—CNPq, together with the Universidade do Vale do Rio dos Sinos. (UNISINOS). The project has a group of 7 partner hospitals, which are: Grupo Hospitalar Conceição[3], Santa Casa de Misericórdia[4], Hospital Ernesto Dornelles[5], Hospital de Clínicas[6], Hospital Moinhos de Vento[7], Hospital Universitário de Santa Maria[8], and Unimed Central de Serviços RS[9].

The MSD project proposes developing an intelligent information and distributed Blockchain-based model architecture with standardized clinical data, allowing to interconnect healthcare service providers and patients. In the COVID-19 pandemic scenario, a disease caused by the new coronavirus called SARS-COV-2 instigated the scientific community regarding innovations and technologies that can improve healthcare services provided globally.

This research aims to contribute to both aspects, observing the need to move forward with structuring unstructured data and ensuring quality over the already collected data. We propose to study the adoption of state-of-the-art IE approaches into non-structured clinical data using NLP tasks such as Named Entity Recognition (NER), structuring data into a semantic model to represent clinical COVID-19 inpatient data.

By proposing this model we aimed to employ Natural Language Processing (NLP) techniques combined with machine learning in a hybrid approach, applied to unstructured textual data from clinical notes provided by partner hospitals. The goal was to structure

[3]Grupo Hospitalar Conceição <https://www.ghc.com.br/>
[4]Santa Casa de Misericórdia <https://santacasa.org.br/>
[5]Hospital Ernesto Dornelles <https://www.hed.com.br/>
[6]Hospital de Clínicas <https://www.hcpa.edu.br/>
[7]Hospital Moinhos de Vento <https://www.hospitalmoinhos.org.br/>
[8]Hospital Universitário de Santa Maria <https://www.gov.br/ebserh/pt-br/hospitais-universitarios/regiao-sul/husm-ufsm>
[9]Unimed Central de Serviços RS <https://www.unimed.coop.br/web/centralrs>

the data in a format that could faithfully represent the information semantically—bringing the intended meaning while we also ensure the semantic interoperability of the data across hospitals. Consequently, our proposal successfully achieved the objective of providing a model that explores the data in layers.

1.1 Research Question

This research aimed to enable different health system providers to utilize large-scale, unstructured real data to contribute to the composition of a patient's electronic medical record. Consequently, the study sought to answer the following research question:

> *How to **extract**, **disambiguate** and **represent** unstructured health data in an interoperability model using **hybrid approaches** combining natural language processing and machine learning techniques?*

1.2 Goals

The general goal of this research is to develop a semantic interoperability model for unstructured healthcare data, transforming it into a structured format that aligns with healthcare standards. We expect, as a result of the model, to provide formatted data suitable for integration into Electronic Health Record (EHR) systems, ensuring interoperability and data consistency across healthcare providers, reinforcing data-driven decision-making and public policy formulation by improving healthcare delivery through Brazil's Unified Health System (SUS) heterogeneity data ecosystem.

1.3 Specific Goals

- Conduct a literature review to identify critical scenarios to achieve semantic interoperability in EHR systems.
- Evaluate a standard that best fits the regional south of Brazil scenario to allow interoperability in healthcare systems.
- Using Machine Learning and Natural Language Processing techniques to structure data from unstructured health records.
- Achieve a structured data level concerned about quality and integrity for use in EHR systems–through the health standard adoption.

1.4 Contributions

In this context, this research aims to contribute to developing a model for semantic interoperability of unstructured data, using recent information extraction tools and a formal structure for representing health data. The experiments were applied to a case study of COVID-19 data, received from partner hospitals of this research. Based on those needs, our contributions to this research we stand out in two lines:

1.4.1 Healthcare contributions:

- Mapping challenges and opportunities related to healthcare and semantic interoperability by structuring information from the real world. The healthcare domain accompanying the patient's life-long cycle and qualified, sensitive data can better provide a continuity of care by consolidating trusted data and information to data-driven decisions and public policy definitions.
- An exploration of real-world regional data problems applied to unstructured healthcare data from a Brazilian perspective, emphasizing the potential of AI applications in Health for inclusion and equity by bringing value to the invaluable data generated in healthcare public and private sectors.

1.4.2 Computing contributions:

- Proposing a semantic model by identifying the components needed for information extraction on non-structured health records to ensure semantic interoperability.
- Providing a data annotation methodology in real-world healthcare data and delivering a small clinical Portuguese dataset focused on a case study of COVID-19.
- An ontology structure based on clinical notes to represent COVID-19 data as intermediate data representation related to patient evolution and controlled vocabularies – addressing patient records and mapping real-world vocab.
- An experimentation on hybrid techniques to prototype a pipeline using information extraction techniques (NLP and ML) by mapping technological needs, extracting essential information, and representing data into an ontology structure, ensuring care at the semantic level.
- Development of a COVID-19 thesaurus to lexical normalization and semantic matching.

- Proposing an ontology aligned to healthcare standards, facilitating the process for harmonization into a selected standard, such as openEHR, HL7 FHIR, and others.

1.5 Organization

Chapter 2 begins by presenting the fundamental concepts related to formalizing EHR systems interoperability, providing an essential foundation for understanding the challenges and solutions in this field. This chapter also offers an overview of commonly used healthcare standards, crucial for achieving semantic interoperability and machine learning scenarios applying to natural language processing. Building on this foundation, Chapter 3 explores the recent studies through a systematic review of semantic interoperability in health records. It highlights key outcomes from recent research on information extraction solutions, particularly in unstructured data. It demonstrates how Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) can play a vital role in addressing these challenges. The chapter looks at how these technologies contribute to the broader goal of semantic interoperability. Following the theoretical review, Chapter 4 introduces our proposed model for achieving semantic interoperability in unstructured health data. This chapter details the designed model to structure and harmonize the extracted information, providing a concrete solution based on the findings from the literature.

In Chapter 5, we describe our methodology for applying the technologies and tools used in our experiments, along with the validation plan for testing and refining the proposed model. In Chapter 6, we present the results and findings of this research, focused on understanding the nature of unstructured data, analyzing information extraction techniques, and providing the overview of the data extraction process to fill out into the developed ontology. Exploring the proposed model for semantic interoperability of data in the healthcare domain. Finally, in Chapter 7, we synthesize the findings drawn from the study and outline the future perspectives to refine the model further, aiming to expand its applicability across various healthcare domains and data types.

2 THEORETICAL BACKGROUND

This Chapter presents research topics involved in this study, as previously presented. We defined that scenario as hybrid once it combines the use of Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) techniques with an ontology approach. Through that, it promotes the structuring of unstructured health records, allowing the adoption of a health standard to promote interoperability at a semantic level. Therefore, we contextualize the definitions of EHR systems and the most known health standards to enable interoperability. Also, we define Machine Learning and Natural Language Processing as subfields of the Artificial Intelligence field, looking for techniques to facilitate information extraction from unstructured data.

2.1 Electronic Health Records

The Electronic Health Record (EHR) system adoption has been a consolidated practice for health information management and a great challenge regarding the heterogeneity and complexity of data since there is no single international standard (FU et al., 2020).

According to ISO18308 (2011), an EHR aims to integrate health records in a processable way, to allow secure storage and communication, using an information model for sharing data with authorized users, and preserving patient information throughout life, efficiently and with quality, with an integrated health system. The structure of an EHR maintains the patient's health status. This format must be digitally processable (ISOTR20514, 2005) to keep patient data securely stored in a repository throughout its lifetime. There are different data formats in EHR systems, such as structured data and non-structured textual data (MARTIN-SANCHEZ; VERSPOOR, 2014; NEGRO-CALDUCH et al., 2021). EHR covers extensive medical histories, including more patients' complete information on potential risk factors.

On the other hand, with the rise of mobile devices and easy access to the internet, the responsibility for providing the information is no longer centralized only on healthcare professionals but also on the patients. In this sense, Personal Health Record (PHR) definition proposes the patient as an active agent to generate and maintain their data. While the EHR definition aims to allow the management and patient data maintenance to be performed exclusively by healthcare professionals and providers, the PHR wants to share this responsibility. Therefore, a PHR has advantages over an EHR system, allowing the patient to provide, maintain and update their information (ROEHRS et al., 2017). Patients can keep their health history in a single view, regardless of where the patient receives a treatment (ROEHRS et al., 2018). A PHR is intended to complement and not replace an EHR, as it allows individual patients' information to be up-to-date and maintained by themselves, in a separate and secure

environment to receive this information (WANG; DOLEZEL, 2016). In this sense, a PHR provides health information captured and reported by individuals (patients) (TANG et al., 2006).

Furthermore, there are other benefits to adopting an electronic health record, such as support the daily provision of care in hospitals and primary care clinics (BLACKMAN-LEES, 2018), patient data reuse, including managing an individual patient's, medical and health services research, and management of health care facilities (MOREIRA et al., 2018), delivering high-quality care and allocating essential resources (SILVERMAN et al., 2021); cost reduction (greater assertiveness when providing care, and avoid repeated procedures), expense control and save time (SHEN; HSIA; HSU, 2021; WANG et al., 2003); greater therapeutic control and less risk of adverse problems (once the healthcare professionals have an entire patient's data access, allowing greater therapeutic assertiveness (FOREMAN et al., 2020).

Adopting an EHR is essential for healthcare providers' management and data exchange among healthcare professionals and institutions. The EHR allows communication between clinicians, nurses, laboratories, and hospitals despite the different existing systems. According to ISO13606 (2019), sharing data between healthcare organizations and healthcare agents must foster the correct data interpretation, with the same precision and meaning adopted from the sender, achieving semantic interoperability. Which means the main objective to adopt an EHR system, besides other benefices, aims to achieve patient data interoperability between healthcare institutions in an integrated way.

2.2 Interoperability

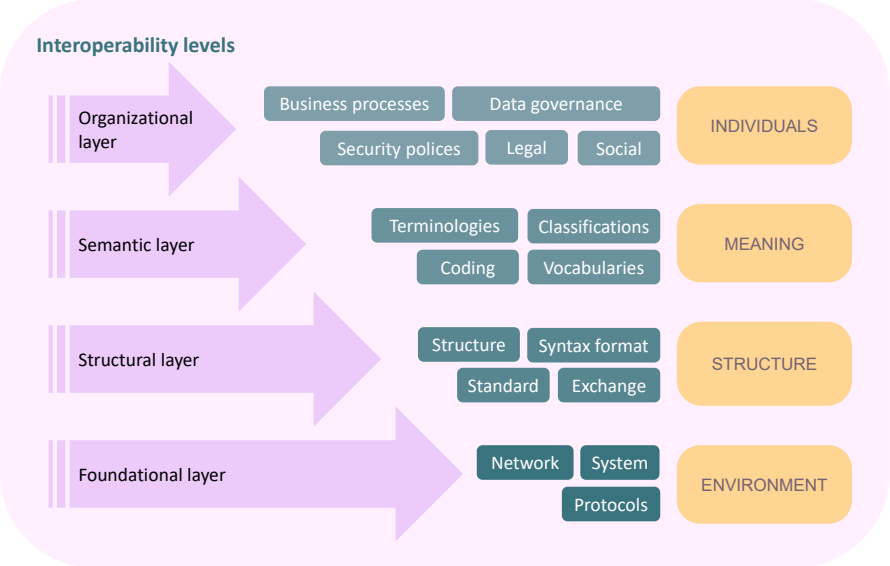
Interoperability is the ability of two or more systems to work together, regardless of interfaces, platforms, and technological choices (BENSON; GRIEVE, 2021). Interoperability focuses on data quality to enable sharing data. Once there are many challenges to sharing data in the healthcare area, such as abbreviations and colloquial expressions, the international health standards lack adoption, controlled vocabularies, and non-structured data (KAH; ZEROUAL, 2021). For example, duplicate entries and errors usually happen because of open textual fields since the systems are not prepared to receive complete information.

According to Sheth (1999), the fundamental problem of data sharing is to maintain consistency when objects move among databases. The HIMSS (2021), a global advisor supporting the transformation of the health ecosystem, with 480 provider organizations and more than 450 non-profit partners – had defined in four levels the interoperability technology: 1) Foundational; 2) Structural, 3) Semantic and 4) Organizational. The **Foundational level** defined requirements to connect different systems, protocols,

networks to collect, store and exchange data. The **Structural level** defined the format, syntax, and data to interpret at the field level; the **Semantic level**–semantic interoperability–aims to share data among organizations or systems and ensure understanding and interpretable data, regardless of who is involved, using controlled vocabularies, terminologies, and formal data representation (BENSON; GRIEVE, 2021).

On the other hand, we can define semantic interoperability separately: semantics study meaning focused on the relationship between people and their words. That is important to help people understand each other despite different experiences or viewpoints (ALEXOPOULOS, 2020). Besides that three levels, some studies discuss the fourth level as essential, the **Organizational**, which regards how the institutions get involved and make better initiatives to the entire healthcare ecosystem inside adopted the proper use of the EHR. Figure 1 represents the four interoperability levels.

Figure 1: The four interoperability levels defined to achieve an EHR interoperable.



Source: Created by the author based on (HIMSS, 2021) .

Therefore, interoperability for healthcare systems involves combining internationally accepted standards, terminology, and record awareness, aiming to generate quality entries by the healthcare professionals team (MALM-NICOLAISEN et al., 2019; NEGRO-CALDUCH et al., 2021). That scenario desires to build the ability to share data between systems and ensure understanding at the domain concept level (ISO18308, 2011).

In this research, when mentioned, we concept terminologies and controlled vocabularies as referred to a set of terms internationally defined to represent a context in the healthcare domain by standardization of terms. According to Silva et al. (2020), it has been a global trend to standardize terms and languages that characterize the

practice of different professions in the healthcare area, as examples SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms), International Classification of Diseases (ICD), Logical Observation Identifiers Names and Codes (LOINC) and others.

2.3 Health Records Interoperability

Despite this scenario of a joint effort among system providers and healthcare providers, there is no global consensus about which healthcare interoperability standards adoption. However, recent studies have proposed some approaches using open standards to achieve interoperable EHR systems (MALDONADO et al., 2020; YANG et al., 2019). The initiative behind the different proposed health standards aims the global adoption of a format for the exchange and sharing of data between institutions, health professionals, and researchers to allow the secondary use of the collected data. In general terms, health standards are rules established to organize the storage format of data collected from health institutions – covering patient data, exams, prescriptions, among other health records (ROEHRS et al., 2018). It defines how to maintain the information, security, privacy concerns, the types of data allowed, and determines what controlled vocabularies the institution uses. Likewise, clinical information models (CIM) often describe the information needed to meet specific contexts, such as outpatient requests, consults, procedures, and medication prescriptions.

According to Adel et al. (2019), Lee, Kim and Lee (2021) openEHR, HL7 and IHE are the most seen standards in studies to propose measures for interoperability, EHR systems, data sharing as well as enabling secondary use, which we present follow.

2.3.1 openEHR

Developed and published by the openEHR Foundation, a non-profit organization, the openEHR standard is an open standard that proposes a multi-level approach to structuring and representing knowledge. Being one of the most internationally recognized standards (KALOGIANNIS et al., 2019), promoting semantic interoperability and service-oriented interfaces (ADEL et al., 2019).

According to the official documentation[1], the openEHR framework is composed of Reference model (RM): a stable reference information model; the first level of modeling **Archetype:** a re-usable content element definitions, to represent the formal definitions from the clinical data; **Template:** a set of context-specific data definitions, used for forms, documents, messages, created by combining required elements of relevant

[1]openEHR Specifications <https://specifications.openehr.org/releases/BASE/latest/architecture_overview.html>

archetypes. In this sense, the RM represents the software implementation. On the other hand, Archetypes define data structures semantically and provide a formal definition to the EHR data, using terminologies and vocabularies to support it.

OpenEHR had been used over the years to achieve semantic interoperability in health records (ELLOUZE; BOUAZIZ; GHORBEL, 2016; MOREIRA et al., 2018; YANG et al., 2019), some recent studies applied combined approaches, using ontologies to represent clinical models (LEGAZ-GARCÍA et al., 2016; LEZCANO; SICILIA; RODRÍGUEZ-SOLANO, 2011; TAWEEL et al., 2011) once the ontological structure is more semantic oriented than the official Archetype Language Definition (ADL) proposed by openEHR-more syntactic oriented.

The openEHR standard has a robust semantic orientation, which makes it easy to use for proper representation of collected data, storage and sharing. On the other hand, recent research has shown interest in using openEHR exclusively for these first two scenarios (representation and storage) and adopting more communication-oriented standards (such as HL7 FHIR) for effective data exchange.

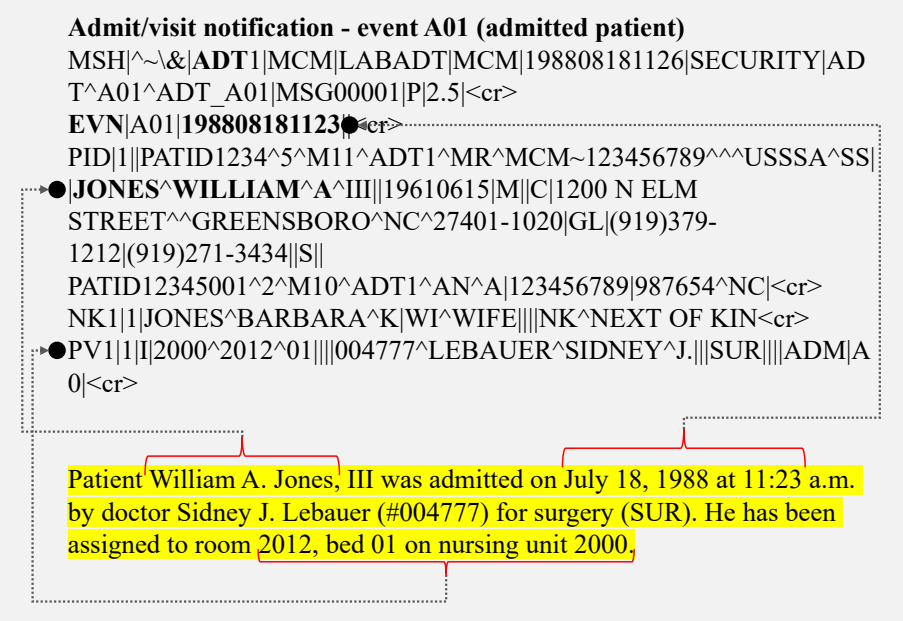
2.3.2 HL7 ecosystem

The HL7 standards were created and maintained by Health Level Seven (HL7), a non-profit ANSI-accredited, responsible for developing standards and promoting a comprehensive set of standards to enable exchange, share and integrate health data. Although we selected the main standards, the HL7 ecosystem has a set of standards and vocabularies[2]. The first HL7 version was released in 1987, known as HL7 V2, and in Figure 2 we show a file snippet of HL7v2 format. The yellow highlighted part in this figure represents the patient information in textual format, and above has an HL7 V2 syntax representing this information.

The PID segment specifies the patient's information, such as gender, address, marital status, citizenship-contains 30 columns with different values separated by the symbols. Table 1 present the primary separators used in the HL7 V2X standard:

[2]Introduction to HL7 Standards <https://www.hl7.org/implement/standards/index.cfm?ref=nav>

Figure 2: An HL7 v2 sample.



Source: Extracted from the HL7 documentation [a]

[a] HL7 Documentation <https://www.hl7.org/>

Table 1: Table to contextualize the type and use of standards.

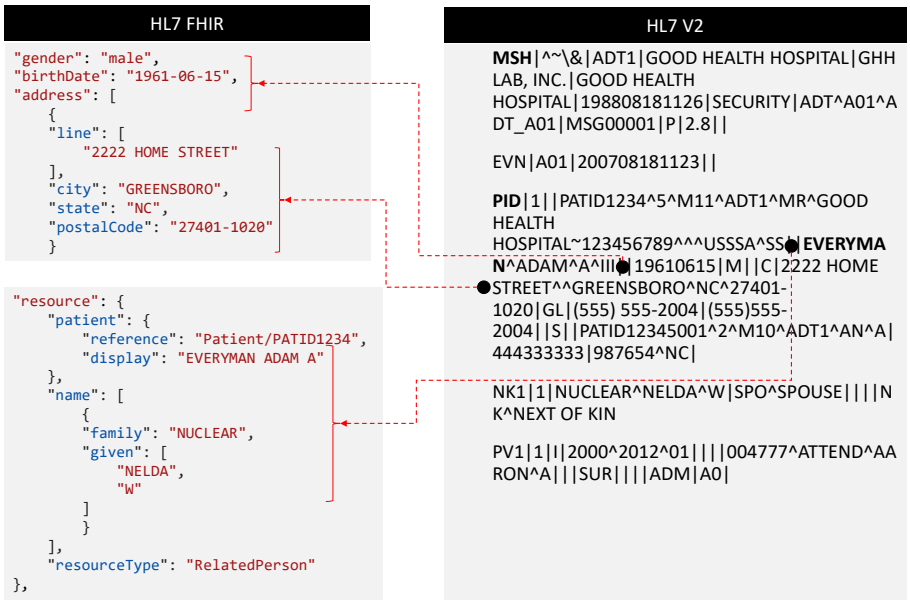
Symbol	Description
<CR>	carriage return – segment separator
(|)	pipe–composite separator
(^)	caret–sub-composite separator
(&)	sub-sub-composite separator
(~)	separate repeating field
(MSH)	header segment
(PID)	patient identity–the 5th field represent the patient's name (last, first, middle name)
(PV1)	information regarding the patient's medical visit

Source: Created by the author.

The **HL7 V2** standard has been in use for the last 25 years, and therefore we can find some initiatives driven by HL7 to convert it to the latest published standard. On the other hand, the newest version released in 2005, **HL7 V3**, sought to solve problems identified in previous versions. However, it is not backward compatibility with HL7 V2X versions. The HL7 V3, unlike HL7 V2X, follows the Reference Infomation Model (RIM)[3]

[3] RIM <http://www.hl7.org/implement/standards/rim.cfm>

Figure 3: Example of HL7 V2 format to HL7 FHIR JSON format.



Source: Created by the author based on HL7 documentation.[a]

[a]HL7 Documentation <https://confluence.hl7.org/display/OO/v2+Sample+Messages>

definition based on ISO/HL7 21731—HL7 V3 Development Framework (HDF). These standards enable data exchange through messages and have a more semantic orientation.

However, when we talk about standards for document representation, HL7 has some standards called 3Cs standards CDA, CCD and CCR, which are: the **Clinical Document Architecture (CDA)** [4] , which specifies the semantics and structure of clinical documents – such as Discharge Summary, Imaging Report, Admission and Physical, Pathology Report, and more. The **Continuity of Care Document (CCD)**[5], built on the CDA standard, which is an implementation guide to implement the **Continuity of Care Record (CCR)**.

Since all the other standards were released before smartphones and digital transformation from personal devices (SARIPALLE, 2019) in 2011, the HL7 published the **FHIR standard** to provide a fresh look to the health standard domain. The Fast Healthcare Interoperable Resources (FHIR) was developed based on the HL7 V3 and HL7 CDA standards but has a leaning focused on web applications. Therefore, the applications that use the FHIR standard are based on RESTFul interfaces (LAMPRINAKOS et al., 2014). Figure 3 demonstrates the format difference between HL7 V2 and HL7 FHIR (JSON) standards. We emphasize that, as HL7 V2 has a necessarily syntactic orientation, it is not possible to represent it in XML or JSON, being only in textual format.

[4]CDA <https://www.hl7.org/implement/standards/product_section.cfm?section=10>
[5]CCD <https://www.hl7.org/implement/standards/product_brief.cfm?product_id=6>

Fundamentally, based on the Resource concept, an FHIR atomic unit, all other FHIR objects are created and accessed through URLs via REST. A set of Resources, built such as blocks aggregating: patient, encounter, condition, procedure, medications, observation, immunizations, and care plan resources (SARIPALLE, 2019). HL7 defines a set of base resources, with the complete health information which can be adapted to as needed to the local requirements through the Profiles (BAE; YI, 2022). The FHIR standard can be defined in JSON or XML format and works with knowledge blocks, Resources published and maintained by HL7. In this way, to use the HL7 FHIR, we define a set of constraints we need, around information to add or remove from the standardized Resources, and generate a Profile from the Resource–resulting in a resource customized to our needs. In the profile, we also define the vocabularies and the kind of data we will collect (integer, string, text, boolean) in compliance with the Resource constraints.

The HL7 FHIR, like multi-level standards approaches, has a specification that facilitates interaction between experts from different fields, separately to healthcare or technology professionals. Essentially, the documentation has five levels, where the first two levels are aimed at technology professionals, while levels 3, 4, and 5 involve health professionals defining clinical information models (CIM). Despite its recent publication, the healthcare community showed good acceptance for adopting the standard, as reinforced by Lee, Kim and Lee (2021), Kasthurirathne et al. (2015), exploring a model to promote EHR interoperability with the proposal to adopt the HL7 FHIR standard for communication between institutions.

In Table 2 we summarize the main characteristics of the described HL7 standards. According to Lee, Kim and Lee (2021), Benson and Grieve (2021), FHIR can improve semantic interoperability since includes the controlled vocabularies and terminologies through the use of CodeSystem and ValueSet definitions. Although it has a complete set of tools for standardizing data to enable interoperability, the HL7 standards do not define criteria for medical imaging - such as computed tomography or ultrasonography, and others. In this sense, the HL7 ecosystem maintains compatibility with the most used medical images standard, the DICOM standard, for managing, storing, and sharing medical images.

2.3.3 DICOM

Aiming to promote a standard that facilitated the interoperability of medical images in medical devices, the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA), though a collaborative work, developed the "Digital Imaging and Communications in Medicine" (DICOM), published in 1985 as ACR-NEMA-v1.0 and later published as DICOM in 1993. The standard enables interfacing between different devices, such as computed tomography, ultrasonography,

Table 2: Table to contextualize the HL7 technical characteristics of standards.

Item	**HL7 V2**	**HL7 V3**	**HL7 CDA**	**HL7 FHIR**
Format	Symbols and characters	XML	XML	XML/JSON/ access from API
Interoperability	Syntactic	Syntactic and Semantic	Syntactic and Semantic	Syntactic and Semantic
DICOM	Limited support	RIM embedded	RIM embedded	RIM embedded
Compatibility	All V2 versions	Not support V2	Not support to earlier version	support V3 and CDA, not support V2

Source: Created by the author.

nuclear medicine scanners (JR et al., 1997). Essentially, DICOM enables the communication and management of medical imaging information and related data (ASSOCIATION et al., 2003) to allow share and store format.

Divide into 20 modules (less two since the update, now 18), the architecture of the standard has sections to define structures and conformance requirements to meet the objective of communication and management of medical information and images (MILDENBERGER; EICHELBERG; MARTIN, 2002). The standard does not just describe the form of an image or data document, like a file. It defines a set of services around the storage, processing, transmission, and electronic format for communication- a whole communication protocol (JR et al., 1997). According to the official documentation of the standard, to meet diagnostic medical imaging from radiology, cardiology, pathology, and others. However, it remains a digital imaging standard and can apply to a wide range of environments(ASSOCIATION et al., 2003). The standard's specifications describe the entire data structure definitions, syntax, data dictionary, message exchange format, web services to exchange, and protocol for communicating the medical images and the related patient data, forms, and reports (GIBAUD, 2008).

Unlike the standards already discussed, DICOM focuses on the specification of standards for working with medical imaging and may eventually contain textual information that relates to medical imaging. For instance, in Herrmann et al. (2018) the authors implemented in the digital pathology environment, since Tissue-based diagnostics remain on images, and pathologists can share with other colleagues and perform a simultaneous review and enable clinical research from other health institutions.

On the other hand, to achieve an interoperable EHR, this standard is supported by other standards intended for textual data specifications. Some initiatives have demonstrated the integrated use of DICOM, as promoted by HL7 through the guidelines defined by the Integrating the Healthcare Enterprise (IHE) and represent an alternative

Table 3: Table to contextualize the type and use of standards.

Name	**Type**
Data format	CSV, XML, JSON, RDF, HTML, TXT
Terminology	LOINC, SNOMED CT, ICD-10
Content EHR	openEHR, ISO 13606, HL7 CDA
Share/Exchange	HL7v2, HL7v3, HL7 FHIR

Source: Created by the author.

to encourage the reuse and share of medical images and medical records EHR. In work by (SLOANE; COOPER; SILVA, 2021), the authors presented an integrated architecture to data collected by medical devices and equipment. The solution promotes a combined approach, integrating IHE and IOT to monitor patients at a distance (observing the pandemic scenario and the need to keep patients safely outside the hospital full of COVID patients), employing HL7 FHIR and DICOM standards. In this line, the authors (MARGHERI et al., 2020) had proposed to adopt HL7 FHIR to exchange data, HL7 to store textual data, and DICOM to support medical image management and their related reports.

Recently, the demand for systems that allow data interoperability at a semantic level has been an object of interest, mainly in healthcare system providers. Different studies explore approaches to solving interoperability problems. However, there are difficulties in adopting health standards and tools to adequate data representation (ontologies, databases, clinical models) that ensure healthcare professionals efficiently manage the data. Figure 3 shows the relationship between usage type and related data format/standard.

2.3.4 Discussion on semantic interoperability

In recent years, the healthcare domain has witnessed a digital transformation across all sectors. Before that, all patient care information interaction stored on paper – this scenario also applies to their management and financial scenarios (SHEN; HSIA; HSU, 2021; WANG et al., 2003).

Electronic Health Record (EHR) systems in healthcare institutions have become a reality. Recent studies have explored new approaches to mediate the EHR adoption (LI et al., 2020b; LEE; KIM; LEE, 2021; MALDONADO et al., 2020; KOPANITSA, 2017), with challenges related to the large volume of data and growing on a large scale (XIAO; CHOI; SUN, 2018; MARTIN-SANCHEZ; VERSPOOR, 2014). As well as the concern with the data collected quality (BLOOMROSEN; BERNER et al., 2020), which often present an unstructured data, or structured format (in complex heterogeneous databases (SIVARETHINAMOHAN; SUJATHA; BISWAS, 2021)), which makes it challenging to adopt healthcare standards. Since that, become infeasible the goals of interoperability,

data reuse, and secondary use of EHR (GAGALOVA et al., 2020) thought a semantic interoperability EHR.

A semantically integrated health system allows sharing data among organizations and their internal ecosystem without missing the meaning. The search for semantic interoperability in health records and different clinical annotations is one of the main challenges for systems, being a constant objective of studies in the last years, such as (MALDONADO et al., 2020; ROEHRS et al., 2019; KOPANITSA, 2017). Implementing interoperability can allow healthcare professionals to manage the complete electronic patient record, regardless of the organization that generated the clinical session entries (MARTÍNEZ-COSTA; MENÁRGUEZ-TORTOSA; FERNÁNDEZ-BREIS, 2010). As reinforced by Kim and Joshi (2021), health record interoperability became a crucial issue in the healthcare scenario, particularly observed during the COVID-19 pandemic to further disease control.

Therefore, recent studies have explored approaches to achieve the interoperability goal, employing semantic web technologies (ontologies) as a support tool (LOZANO-RUBÍ et al., 2016; SINACI; ERTURKMEN, 2013), for data representation (YUKSEL et al., 2016; MENDES et al., 2012) and clinical information models (LEGAZ-GARCÍA et al., 2016; TAWEEL et al., 2011). As well as exploring works in the Natural Language Processing (NLP) and Machine Learning (ML) areas (SHEN; HSIA; HSU, 2021) to collaborate in the data structuring process–enabling an interface between the human language (non-structured data) to a computer machine-readable format (HOUSSEIN; MOHAMED; ALI, 2021).

2.4 Ontology

Since the EHR adoption has different health standards involved to enable data sharing among healthcare organizations, some studies have explored ontology structures to represent data as an intermediary data representation (BLACKMAN-LEES, 2018). However, the adoption of standards still presents several challenges to achieving interoperability at the semantic level. The exponential growth volume data generated in the healthcare domain, one of the reasons being the computerization of hospitals, caused the massive search by better management systems and electronic records (LEE; KIM; LEE, 2021; ADEL et al., 2019).

In this scenario, the need for systems that allow information sharing between institutions is fundamental (YUKSEL et al., 2016) and ontologies had promoting mechanisms to provide integrations in the patient registry ecosystem. Besides, some works develop approaches using ontologies to rules representation and constraints from Clinical Information Models (CIM) (LEGAZ-GARCÍA et al., 2016; LEZCANO; SICILIA; RODRÍGUEZ-SOLANO, 2011; TAWEEL et al., 2011), representation of data

extracted from EHR (LEZCANO; SICILIA; RODRÍGUEZ-SOLANO, 2011; MENDES et al., 2012), as well as in employment as an intermediate object, containing the interaction rules, terminologies, classifications (ELLOUZE; BOUAZIZ; GHORBEL, 2016; MARTÍNEZ-COSTA; MENÁRGUEZ-TORTOSA; FERNÁNDEZ-BREIS, 2010; LEGAZ-GARCÍA et al., 2015; MARTÍNEZ-COSTA et al., 2015; YUKSEL et al., 2016).

Moreover, the authors Hitzler (2021) had highlighted the semantic web to share data, integrate, and reuse through ontologies, linked open data, and knowledge graphs to ensure the correct meaning of shared data. The semantic web is also known as one of the fundamental technologies to achieve semantic interoperability in health information systems Demski, Garde and Hildebrand (2016), often using ontologies, a well-established technology to support knowledge-intensive tasks related to EHR systems Maldonado et al. (2020).

In order to allow the sharing and reuse of formally represented knowledge between systems, the author Gruber (1993) proposed a representational vocabulary specification in order to share a domain of discourse, through classes, relations, and functions. According to Gruber (1993), the main target to adopt ontologies is to share a common understanding of the information structure among people or systems. Ontologies allow a formal and structured knowledge representation, enabling the reuse and data sharing (RUIZ-MARTÍNEZ et al., 2011).

The term ontology is defined by Gruber (GRUBER, 1993) as an explicit, formal specification of a shared conceptualization, with controlled vocabulary to describe the main concepts and relationships about a specific domain. Conceptualization represents a body of information formally expressed (GRUBER, 1995), the act of linking the relating concepts to the interest domains, as well as the relationships that exist between these (ABDELGHANY; DARWISH; HEFNI, 2019).

Furthermore, the health domain has benefited by adopting ontologies as a representation format for controlled vocabularies in the area (NOY et al., 2009; SIVARETHINAMOHAN; SUJATHA; BISWAS, 2021). Ontologies are also regarded as the foundation of knowledge representation and reasoning There are numerous ontologies already published for the health field today and available for access on the website BioPortal[6].

2.5 Natural Language Processing

Considered a field of Artificial Intelligence area, Natural Language Processing (NLP) (MANNING; SCHUTZE, 1999) origins at the intersection of computation and linguistics. The interdisciplinary research field is concerned with interfaces between human languages and computers (SHEN; HSIA; HSU, 2021), as it refers to the objective of making computer

[6]BioPortal <https://bioportal.bioontology.org/>

systems able to analyze, understand and produce human languages in their different languages, such as English, Spanish, Portuguese, German and others (ALLEN, 2003).

The popularity of the area can be related to the rapid growth of data available in unprecedented volumes, a format usually is unprepared for processing by machines, and NLP has been approaching the areas of Big Data and Machine Learning (MARTIN-SANCHEZ; VERSPOOR, 2014) to contribute and to structure these data. Furthermore, the study conducted by Xiao, Choi and Sun (2018), Martin-Sanchez and Verspoor (2014) highlighted the healthcare domain as a rising research area for the NLP, Machine Learning (ML), and Deep Learning (DL) areas because of the large volume of unstructured data generated daily.

The exponential growth of unstructured data with the birth of the internet has made NLP one of the essential technology areas (HASAN; FARRI, 2019), which also causes a digital transformation in the healthcare domain by the EHR adoption. The healthcare institutions began to migrate their patients' history of paper records to electronic records, making the area a target to research. The unstructured data reality has exacerbated this difficulty Martin-Sanchez and Verspoor (2014), furthermore, combined with the adoption of the personal devices to support patient monitoring activities (HEYMANN; SHINDO, 2020), generating more information within health institutions, high demand for unstructured data. In addition, these data come from different sectors (e.g., laboratory exams, patient data, medications, prescriptions, clinical notes, medical images, and related reports) (NEGRO-CALDUCH et al., 2021), generating fragmented databases (heterogeneous data that need to be related).

In this sense, Information Extraction (IE) in EHR systems has become a necessity and a rapidly developing area. Apply techniques involving ML and NLP in order to enable systems to understand syntactic structures in free text, as well as extract the meaning behind a narrative (DASH et al., 2019). The NLP Information Extraction (IE) task focuses its efforts on extracting implicit and explicit information (OTTER; MEDINA; KALITA, 2020). While the first relates to words and linguistic structure in general, as discussed earlier, the second (explicit information) means information obtained through semantic relationship and context. In the NLP area, the main tasks dedicated to this activity and information extraction are: Named Entity Recognition, Relation Extraction, and Entity Linking.

There is no exact distinction between unstructured and structured data. The first, in general, can present some structure such as metadata or attached documents. In contrast, structured data can present heterogeneous bases and tabular data, making it difficult to contextualize the patient's global scenario since there are different bases (MARTIN-SANCHEZ; VERSPOOR, 2014). While the unstructured format presents the biggest obstacle in information extraction to enable the use by clinicians, nurses, and secondary use (HASSANPOUR; LANGLOTZ, 2016), it also represents excellent

opportunities, once bring a complete narrative from different professionals into the patient record-even unstructured format (SILVERMAN et al., 2021). Therefore, recent studies (LI et al., 2021; LI et al., 2020a) have discussed promising approaches combining approaches in the areas of NLP and ML to apply information extraction techniques.

Named Entity Recognition (NER, also called entity identification or entity extraction) is the first step for information and knowledge acquisition when we deal with unstructured texts (RIVERA-ZAVALA; MARTÍNEZ, 2021). Named Entity Recognition aims to classify text into recognized known entities. This process happens since there are previously annotated data, which allows the recognition of these categories. The annotated data contain the entity and its respective concept (meaning) information. We can say that NER seeks to locate and classify text in natural language through previously categorized entities. According to Li et al. (2020a), some examples of named entities are organization, person, and location names in the general domain. However, in the Biomedical Domain, often adopted the term Biomedical Named Entity Recognition (BioNER), which aims to extract specific entities (e.g., gene, protein, drug, and disease names in a biomedical domain) (NASEEM et al., 2020).

Relation Extraction (RE) attempts to fill this gap by extracting semantic relationships between entity pairs from plain text. This task can be modeled as a simple classification problem after the entity pairs are specified Vashishth et al. (2018). The task of relation extraction (RE) aims to identify semantic relationship between entities or entity pairs **e1** and **e2** in a given sentence Gupta et al. (2019), Otter, Medina and Kalita (2020). According to Konstantinova (2014), RE aims to discover entity pairs through the different semantic connections between the recognized entities, the reason why NER and RE are typically employed together.

As explained by Nasar, Jaffry and Malik (2021), in the Relation extraction task, we can separate some different methods: Rule-based, applying a sequence of rules defined manually to identify patterns from the analyzed text and extract the relations (e.g., adopt regular expression to extract relations); Semi Supervised-based, the first rules are defined based on initial patterns, and the following new rules are created automatically from the unlabeled text data; Supervised-based, usually run a classifier to train in an annotated data which employ some approximation approach like proximity distance between words, dependency path. Distance Supervision-based uses the supervised method, but the annotations come from a knowledge base, then classifier the relation.

In the information extraction (IE) scenario, around the Named Entity Recognition and Relation Extraction tasks, often occur problems with entities representing more than one concept. According to Newman-Griffis et al. (2021), ambiguity consists of the scenario when there are words and phrases with more than one concept - a usual problem solved jointly with extracting information from the biomedical domain.

The Named Entity Disambiguation task, also called Entity Linking, according to Ji,

Wei and Xu (2020) involves three challenges. The entity recognized successfully received more than one concept, resulting in an **ambiguous** term, since there is not a single context established. However, the opposite can also happen, a concept (context) linked to many entities, known as **variation**. On the other hand, if there is no entity recognized, we define it as **absence**, which can happen for several reasons, such as the term being rare, the knowledge base used not being able to link the term to context. Furthermore, the NED task can be confused with WSD (Word Sentence Disambiguation). Despite sharing the objective, the two employ different approaches. As highlighted by Moro, Raganato and Navigli (2014), NED essentially focuses on generating candidate entities and evaluating which of these best fit the objective, often using an open knowledge base. In contrast, the WSD task is concerned with connecting the recognized entity with the context (entity-concept) through exhaustive dictionaries from the specific target domain. In the medical domain, Entity Normalization has a particular application since the recognized entities need to be related to controlled vocabularies Newman-Griffis et al. (2021). The disambiguation tasks aim to link recognized entities to the respective vocabularies. (VRETINARIS et al., 2021). The biomedical domain ofter uses acronyms, and non-standard clinical terminology by health care professionals in the non-structured data also becomes difficult for the de-identification and anonymization goals on the non-structured text (HOUSSEIN; MOHAMED; ALI, 2021). The study conducted by Bouarroudj, Boufaida and Bellatreche (2022) also reinforces the good results of NED for the disambiguation task, employing resources to calculate the similarity between the relations of the entities to generate the candidates. On the other hand, the authors Li et al. (2019) explored the BERT architecture, using the BioBERT, and ran a fine-tune with EHR notes, to improve the results around the target health domain-showing promising results. The BioBERT model is trained on PubMed scientific publications and, because of that, has a health domain context. However, using EHR notes, the authors are concerned with specific terms.

2.6 Machine Learning

The definition of the term Machine Learning is closely linked to the first steps in Artificial Intelligence (AI), trying to emulate the human form to learn from the information they obtain from the real world and how they process this information to achieve a goal (NAQA; MURPHY, 2015). The access to computers has brought another perspective on what we hope machines can automatically learn and understand (MITCHELL, 1997b). Essentially, when we talk about emulating the human way of processing information, it is the extremely connected form of the human brain, which is "emulated" by neural networks, composed of interconnected nodes (neurons), where

each one receives input weights to generate output data (OTTER; MEDINA; KALITA, 2020).

This “learning” happens from a large volume of data as examples, which it will use in the training process, which means that learning is the result of learning from the experience from data—the data of the target domain (AGRAWAL et al., 2020). Furthermore, there are different kinds of "learning" methods; it can happen through some approaches (GOODFELLOW; BENGIO; COURVILLE, 2016). The **Unsupervised** learning method aims to identify patterns within a large volume available of data, which has no labels, thus obtaining the structure and patterns of a generalized–learning from x distribution features. In contrast to this method, when we run **Supervised** learning method, it applies a set of previously annotated data, which provides the necessary knowledge for the algorithm to identify, from the sample, how it should categorize the data. That is the most common learning approach (LECUN; BENGIO; HINTON, 2015).

On the other hand, the **Semi-supervised** learning method partially applies both methods. The algorithm uses a large volume of data without training annotation without identification; and a small annotated dataset. The authors (BEAULIEU-JONES; GREENE et al., 2016) propose the use of this approach in a scenario where the data with patient information represents the dataset without labels – performing the unsupervised learning, and a small set of electronic phenotyping annotated data – performing the supervised learning. Furthermore, **Reinforcement** learning learns through a reward system. Depending on the performance and the outcome received, the answer (reinforced) will be a positive or negative response, learning through the interaction (NAQA; MURPHY, 2015).

Although, in practical terms, it has similar traits to the machine learning application, deep learning is a neural network with deeper layers. Learning methods at various levels of representation, with multiple simple and non-linear modules. (LECUN; BENGIO; HINTON, 2015). Each module has a transformation per level – the initial input receives raw data and, at the end, with an abstract representation data.

The ML applications explore many areas, such as computer vision, speech recognition, natural language processing, and robot control. (JORDAN; MITCHELL, 2015). Even in its first years, applying methods for extracting information from medical records stood out as an emerging research field. In the article “Does Machine Learning Work?” Tom M. Mitchel presents the health domain as a potential and promising area in which to apply ML since there is an extensive demand for data. The medical domain generates data in many formats and media, such as images (sonograms and x-rays) and text (notes, reports, and patient records) (MITCHELL, 1997a).

In this scenario, recent works corroborate this prediction, applying different ML and DP techniques combined with NLP techniques to promote information extraction in

medical domain data (NEGRO-CALDUCH et al., 2021; HOUSSEIN; MOHAMED; ALI, 2021; OTTER; MEDINA; KALITA, 2020; HASAN; FARRI, 2019). The review conducted by the authors (HOUSSEIN; MOHAMED; ALI, 2021) highlights the RNN architectures (i.e., LSTM, BiLSTM, GRU) and CNN, while the authors Negro-Calduch et al. (2021) reinforce the trend in the health domain, the use of NLP techniques combined with DL and use of ontologies.

Furthermore, the most adapted to disease classification and detection tasks in Deep Learning involves LSTM architecture and Transfer Learning (TL). The authors Laparra et al. (2021) reinforced the use of Transfer Learning to perform NLP tasks to process EHR data, adopting Bi-LSTM, Bi-LSTM-CRF, CNN, and BERT architectures, supporting the need and difficulty to take annotated data. Furthermore, the authors Li et al. (2021) analyze the scenario of unstructured data in the health domain and reinforce the trend of information extraction tasks adopting the RNN (Bi-LSTM and LSTM), BERT (BioBERT), and CNN architectures.

In the health domain, annotating data tend to take a considerable amount of time, one of the reasons for the recent trend for new approaches to adopt unsupervised learning or less annotated data needed. Thus, noting this lack of annotated data, Radford et al. (2018) had proposed a pre-trained approach, called Generative Pre-Training (GPT), which can be considered the basis for the most recent architectures of pre-trained models. Later that, followed by the adoption of a bidirectional approach to expand the context knowledge –forward and backward captured, where we can see the proposal of Embeddings from Language Models (ELMo) (NEUMANN et al., 2018).

Bidirectional Encoder Representations from Transformers (BERT)

Transformer architecture has been widely explored in a variety of activities, being one of the ML approaches with significant results in the health domain for extracting information from medical data (BENGIO; LECUN; HINTON, 2021). Regarding the Bidirectional concept (forward and backward capture), the authors Devlin et al. (2018) proposed the transformers (PENG; YAN; LU, 2019) approach that brings the unsupervised bidirectional pre-training mechanism, combining two unsupervised steps: Masked Language Model (MLM) and Next Sentence Prediction (NSP), and later you can fine-tune the task of interest using annotated data (ZHANG et al., 2020).

In this scenario, the pre-trained proposal allows models to be trained on generic corpus and subsequently easily adapted to specific tasks through the transfer learning techniques (LAPARRA et al., 2021). Another benefit of the BERT model also resides in applying self-attention, which allows capturing long-distance relationships in the (LUOMA; PYYSALO, 2020) input sentences.

The healthcare domain has represented an attractive area for new research. There is a growth in EHR systems adoption, which causes a lack of adequate tools to collect, manage, and share health information. In this sense, researches aiming to propose ways

to intermediate to exchange data among institutions had been proposed (MALDONADO et al., 2020; LEE; KIM; LEE, 2021), besides promoting the healthcare standard adoption as the principal path to achieve interoperability at the semantic level to enable sharing data. On the other hand, the non-structured data volume reality inside the healthcare institutions and the unprepared systems to attend that healthcare records demand (NEGRO-CALDUCH et al., 2021; MARTIN-SANCHEZ; VERSPOOR, 2014) had provoked a growing interest from other research areas. We highlighted the natural language processing initiatives to promote information extraction tasks, combining machine learning approaches (XIAO; CHOI; SUN, 2018). Also, as reinforced by (BLACKMAN-LEES, 2018; ADEL et al., 2019), using ontology as the intermediary structure has demonstrated a promising approach once there is not a consensus about what the best healthcare standard adopts.

In this scenario, regarding the growing interest and EHR adoption in health institutions, next chapter, we present the systematic review conduction to map the main approaches to achieve semantic interoperability in electronic health records. Moreover, the technological efforts involved structuring, information extraction, and data standardization.

3 RELATED WORK

Semantic interoperability in electronic health records (EHR) is challenging, primarily due to the data heterogeneity of healthcare provider systems. Adopting healthcare standards has proven successful in enabling interoperability among different systems and facilitating data exchange. In this chapter, we focused on identifying related works that address the gaps in existing models, technologies, and approach proposals following recent research to achieve semantic interoperability in health records.

First, to perform a literature review, we defined a methodology to conduct the systematic literature review (SLR) focused on semantic interoperability in health records; we discussed the recent articles exploring semantic solutions in non-structured data; in the end, we explored, as final remarks, how we could improve semantics in the healthcare interoperability domain.

Our literature review related to semantic interoperability in electronic health records, as a result of this chapter and research, was published in the Health and Technology journal: *de Mello, B.H., Rigo, S.J., da Costa, C.A. et al.* ***Semantic interoperability in health records standards: a systematic literature review****. Health Technol. (2022). https://doi.org/10.1007/s12553-022-00639-w*

3.1 Systematic literature review

This section describes the protocol we used in the literature review and shows the research questions we designed to extract our specific information of interest from the studies. The research strategy for filtering and selecting the studies involved the adoption of Inclusion and Exclusion criteria. Then, a filtering process was applied to the studies related to semantic interoperability and quality assessment, which was then later applied to answer the research questions.

3.1.1 Study design protocol

A systematic literature review is a research method that allows the identification, evaluation, and interpretation of the studies without bias in the process (ROEHRS et al., 2017). The protocol followed the steps from previous works of Kitchenham and Brereton (2013), aiming to map relevant and recent research in semantic interoperability.

- **Research questions:** introduce the research questions investigated;
- **Search strategy:** outline the strategy and libraries explored to collect data;
- **Article selection:** explain the criteria for selecting the studies;

Table 4: Research questions to extract information of interest from the articles selected in this review.

Group identifier	**Issue**
General questions (GQ)	
GQ1	What is the state of the art in health standards applied in health records?
GQ2	What are the challenge and open questions to semantic interoperability in health records?
Specific questions (SQ)	
SQ1	What are the health standards adopted in the studies?
SQ2	What are terminologies or health repositories used?
SQ3	What are the approaches used?
SQ4	What are the main security concerns used?
SQ5	What are the evaluation approaches used?

Source: Created by the author.

- **Quality assessment:** describe the quality assessment applied to the selected studies;
- **Data extraction:** compare the selected studies and research questions.

We conducted a systematic literature review focusing on semantic interoperability in health records. The studies accepted in this review filled the requirements from the defined protocol and followed the main keywords as below:

3.1.1.1 Research questions

The research questions represent a fundamental part of the systematic review Kitchenham and Brereton (2013), as they allow directing the research and extracting information as needed, as they follow the topic of interest. Table 4 presents the research questions of interest in the systematic review. There are two sets of questions; the first is global context questions, which include an overview of all studies. The latter are specific questions to answer about each article.

3.1.1.2 Search strategy

The search strategy aims to find studies to answer the research questions through the definition of keywords and the scope of the research. The keywords defined focus of interest, which ensures an accuracy level in the search results. The correct selection of keywords ensures adequate research results on the databases Kitchenham and Brereton (2013). We used the five PICOC questions (Population, Intervention, Comparison, Outcome, and Context) to define the scope of the research, as explained below.

Table 5: The final search string defined to search on the chosen databases.

"semantic interoperability" AND ("health record" OR “medical record” OR “patient record” OR “hospital record”) AND standard

Source: Created by the author.

The PICOC acronym **Population** involves keywords, related terms, and variants around the interest area. We used “semantic interoperability” as the central term in the search string shown in Frame 1. In the **Intervention**, being “semantic interoperability” a general term, we used some other keywords to filter: health record, medical record, patient record, and hospital record, aiming to allow filtering semantic interoperability solutions applied in the health context, especially in health records. **Comparison**, this research aims to find different health standards allowing semantic interoperability in health records. The **Outcomes** determine relevant studies that can answer the research questions. This research focused on studies proposing solutions to semantic interoperability problems in health records.

Also, we evaluated studies that explore the limitations of applying health standards. We define **Context** as the concern to identify the different application scenarios and data used on the proposal. This research focused on studies proposing solutions to semantic interoperability problems in health records. In addition, we notice that different kinds of data come from the real world, and most come from a controlled environment (collected straight to the evaluation) or simulated data.

The research string formatted in Table 5 demonstrates the mandatory terms, such as semantic interoperability and their acronyms, health record and their acronyms, and the term standard. At the selection step, the terms applied aim to filter the articles by title, abstract, and keywords.

3.1.1.3 Article selection

Next, the studies with neither a key term addressed nor a PICOC definition aligned, called impurities, are removed. Finally, qualify our results, and for that, we defined the exclusion criteria based on the research question:

a) **Exclusion criterion 1:** the article is not in the range (search date 2010 to September 24, 2020).

b) **Exclusion criterion 2:** the article does not address semantic interoperability or related acronyms (population criteria I).

c) **Exclusion criterion 3:** the article does not address the health record, medical record, patient record, hospital record, and related acronyms (intervention criteria II).

d) **Exclusion criterion 4:** the article does not address standard or related acronyms (comparison criteria III).

e) **Exclusion criterion 5:** the article has less than six pages.

f) **Exclusion criterion 6:** impurities articles (e.g., duplicate papers and non-English studies).

3.1.1.4 Quality assessment

The quality assessment consists of a filter focused on the quality and relevance of studies related to the topic of interest. The quality assessment ensures selected articles have a relevant impact factor in the research area. As a criterion to filter the rest studies, we defined the h5-index score, equal to or higher than 28. If the article passed the cutoff, then it can be accepted in this review.

3.1.1.5 Data extraction

The data extraction step occurs after finishing the selection of articles. This means satisfying the inclusion and exclusion criteria and filtering by the quality assessment. The other articles represent the corpus qualified for full reading and analysis. Survey questions act as targeting filters to extract the information from selected studies according to the topic of interest. Data extraction aims to correlate the area of interest with the defined research questions. The studies propose different solutions to the research problem presented: what health standards approaches can solve semantic interoperability problems in health records? The questions have a limited response context to avoid bias, to respond objectively.

The methodology to answer the research questions proposes to extract the results as presented in Table 6, where the **General Question (GQ)** and **Specific Question (SQ)** have an expected location to search for data.

3.1.2 Results and discussion

Through the SLR, we caught the selected studies concerned with solving interoperability problems. As we intend to identify health standards and technologies usually applied, we opted to search in databases covering both health and technology, such as ACM Digital Library, IEEE Xplore Library, Science Direct, Springer Link, Web of Science, Medline PubMed. We knew Google Scholar would bring too many duplicated results. Still, to ensure we would lose any essential studies, we include and remove the repeated studies from Google Scholar and keep the original publisher journal.

Table 6: Follows we shows the quality assessment of article and related questions.

Section	Description	Research questions
Open Content		
-	Title	GQ1, GQ2, SQ1
-	Abstract	All questions
-	Keywords	GQ1, SQ1
Article Content		
-	Introduction	All questions
-	Method	All questions
-	Results	All questions
-	Evaluation	All questions
-	Conclusion	All questions

Source: Created by the author.

The search step in the databases, as mentioned above, aimed to index the search for studies published in the last ten years. Each database presents a way of formatting the survey, which we respect and modify to suit, but we kept all the mandatory terms defined in the PICOC strategy. Finally, after applying the queries to the search bases, we had a total of 6,032 articles. The initial filter aims to remove patents and citations, non-English studies, which resulted in around 783, roughly because some patents also appeared as citations.

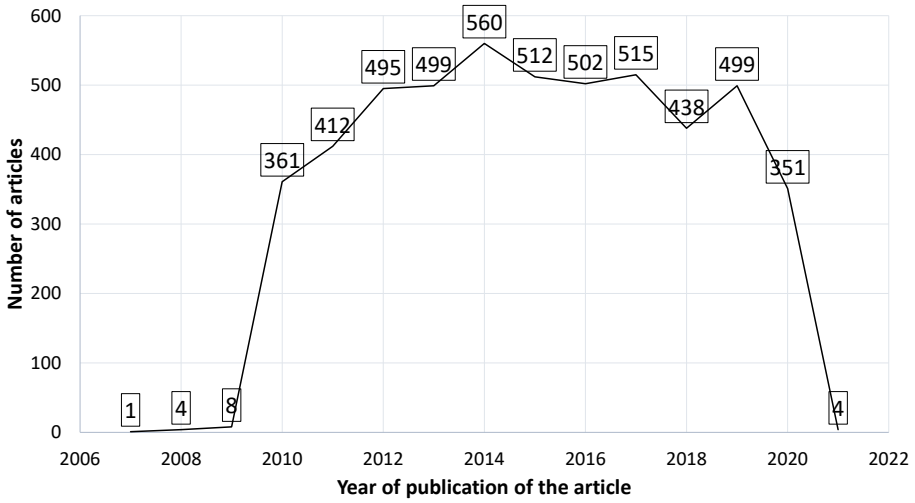
As shown in Figure 4, by year, the published articles in this area have been an interest constant in recent years. For this systematic review, the cutoff was September 22nd, 2020.

3.1.3 Article selection

Figure 5 shows the selection steps, removing impurity studies unrelated to the area of interest. Usually, these impurities studies had references related to the research area or citations of related works, however, without directly informing the area of interest. In addition, articles with less than six pages and no abstract, about 735 impurities, were excluded. We removed non-primary studies, such as editorials, chapters, thesis, reviews, and reports, approximately 1195 studies.

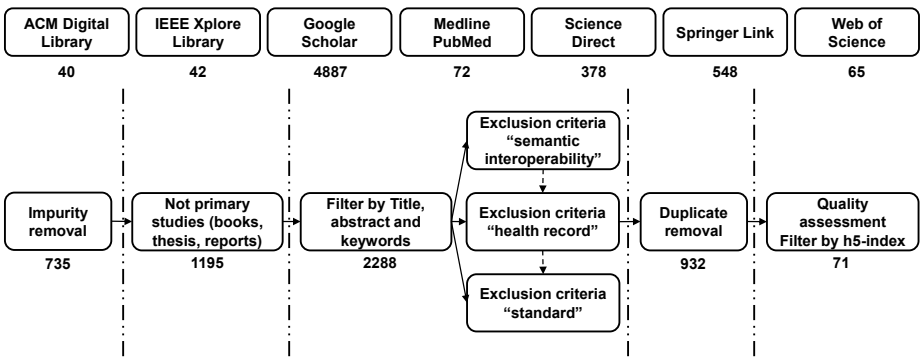
Then, we applied the inclusion and exclusion criteria based on the PICOC strategy: semantic interoperability, medical record (variants such as health record, medical record, patient record, hospital record), and standard. Finally, the inclusion and exclusion criteria are applied to the remaining studies to filter articles directly related to the research topic, totaling 2,288 excluded studies. The filters allow selecting studies according to defined terms.

Figure 4: This graph presents, distributed by year of publication, the corpus of articles published between 2010 to September 2020 found through searches in the selected research bases.



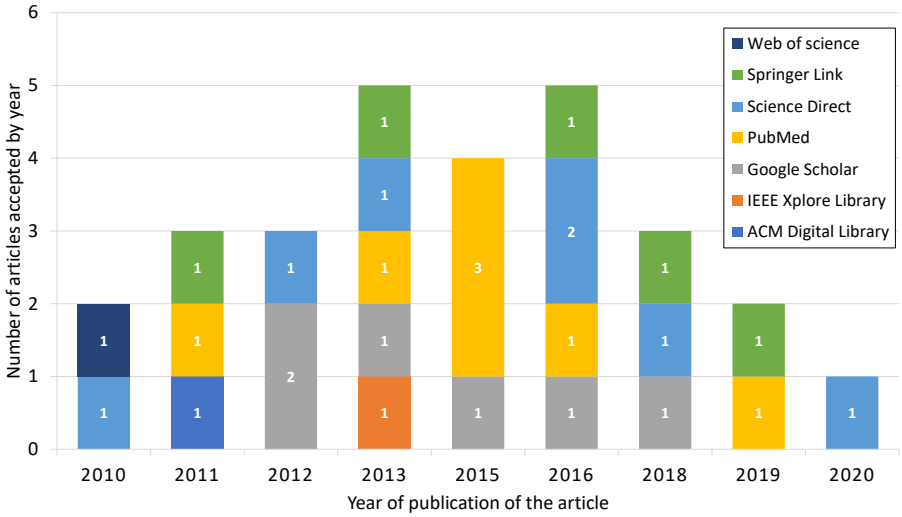
Source: Created by the author.

Figure 5: This figure presents the entire selection process of the studies across the inclusion/exclusion criteria and quality assessment to conduct this systematic review.



Source: Created by the author.

Figure 6: This graph presents the articles accepted after the selection process, showing the number of articles by published year.



Source: Created by the author.

3.1.4 Quality Assessment

The last step had two parts; first, we filtered the studies by main interest area and evaluated the remaining corpus about its objectives. Many studies satisfied the inclusion and exclusion criteria, but they differed from the review's interest topics. Thus, the quality assessment, the last step of the protocol, provides a cutoff parameter, where we look for studies published in relevant journals for the area of interest. Among these, we selected the highlighted studies. For this, we use the h5-index metric, which quantifies the relevance of newspapers and conferences in the last five years, a [1]Google scholar metrics and works with the highest number H. Some studies that presented relevant discussions but did not appear in this index were separated to contribute to our discussion section.

In the final selection, some bases showed a predominance of accepted articles. For example, **ScienceDirect** with eight studies followed for **PubMed Medline** with seven and **Google Scholar** with six studies. Next was **SpringerLink** with four and I**EEE Explore** with two, while **ACM Digital Library** and **Web of Science** had one from each research base. Figure 6 present the final corpus articles accepted in this systematic review.

Figure 6 shows the accepted articles distributed by year, and Table 7 shows which research databases it is from (journal or congress), the number of articles per location, and the H5 index used in the quality assessment stage. Three times the journals had an above-average acceptance, a scenario explained by the applied area, which is related to the research question of this review. We highlight the International Journal of Medical

[1]Google scholar metrics <https://scholar.google.com/citations?view_op=metrics_intro>

Table 7: This table shows the number of papers by year of publication after finishing the selection and filtering steps.

Journal / Conference	**H5-index**	**Amount**
ACM International Conference on Information and Knowledge Management	54	1
BioMed Research International	97	1
BMC Medical Informatics and Decision Making	41	4
BMC Research Notes	44	1
Computer Methods and Programs in Biomedicine	55	2
Future Generation Computer Systems	86	1
Hawaii International Conference on System Sciences	45	1
IEEE Journal of Biomedical and Health Informatics	69	1
International Conference on Autonomous Agents and Multi-Agent Systems	40	1
International Journal of Medical Informatics	55	2
Journal of Biomedical Informatics	60	5
Journal of Medical Internet Research	96	1
Journal of Medical Systems	58	1
Journal of the American Medical Informatics Association	67	4
Knowledge-Based Systems	85	1
Procedia Technology	34	1

Source: Created by the author.

Informatics, with five accepted studies, BMC Medical Informatics and Decision Making, and the Journal of the American Medical Informatics Association, with four accepted studies each one, the last publishers had less than that, one and two, as shown below.

3.1.5 Data extraction

The information extracted from the articles aims to answer the question of interest in the review and identify approaches and solutions that allow achieving semantic interoperability in health records. Table 8 presents the final list of articles selected for this systematic review.

Follows the analysis of the selected studies against the questions of interest, and each other answered individually.

3.1.5.1 SQ1 – What are the health standards adopted in the studies?

There is no consensus on a global standard for electronic health records, and the studies selected for this review reinforced this scenario. However, it is possible to observe a trend in the standard choice with a multilevel approach, as such openEHR, ISO/CEN 13606, and HL7 formats. That dual model approach allows specialists in health and technology

Table 8: This table presents the first author and respective studies by year, relating the publisher and the kind of place it was published.

Identifier	First author	Publisher	Type
A01	Bahga and Madisetti (2013)	IEEE	Journal
A02	Blackman-Lees (2018)	IEEE	Conference
A03	Marcos et al. (2015)	Oxford Academic	Journal
A04	Sáez et al. (2013)	Elsevier	Journal
A05	Legaz-García et al. (2015)	Oxford Academic	Journal
A06	Legaz-García et al. (2016)	Elsevier	Journal
A07	Bono et al. (2011)	BioMed Central	Journal
A08	Demski, Garde and Hildebrand (2016)	BioMed Central	Journal
A09	Duftschmid, Wrba and Rinner (2010)	Elsevier	Journal
A10	Duftschmid, Chaloupka and Rinner (2013)	BioMed Central	Journal
A11	Ellouze, Bouaziz and Ghorbel (2016)	Elsevier	Journal
A12	Fernández-Breis et al. (2013)	Oxford Academic	Journal
A13	Kalogiannis et al. (2019)	BioMed Central	Journal
A14	Lezcano, Sicilia and Rodríguez-Solano (2011)	Elsevier	Journal
A15	Min et al. (2018)	BioMed Central	Journal
A16	Maldonado et al. (2020)	Elsevier	Journal
A17	Marco-Ruiz et al. (2015)	Elsevier	Journal
A18	Moreira et al. (2018)	Elsevier	Journal
A19	Martínez-Costa, Menárguez-Tortosa and Fernández-Breis (2010)	Elsevier	Journal
A20	Martínez-Costa et al. (2015)	Oxford Academic	Journal
A21	Menárguez-Tortosa, Martínez-Costa and Fernández-Breis (2012)	Springer	Journal
A22	Mendes et al. (2012)	Elsevier	Conference
A23	Yuksel et al. (2016)	Hindawi	Journal
A24	Lozano-Rubí et al. (2016)	Elsevier	Journal
A25	Sinaci and Erturkmen (2013)	Elsevier	Journal
A26	Taweel et al. (2011)	ACM	Journal
A27	Urovi et al. (2012)	IFAMAS	Conference
A28	Yang et al. (2019)	JMIR	Journal

Source: Created by the author.

Table 9: Health standards used in the selected studies.

Health standard	**Reference articles**
openEHR	A18, A28, A01, 16, A06, A11, A19, A15, A08, A20, A10, A14, A05, A17
ISO 13606 [a]	A24, A16, A06, A19, A09, A21
HL7 CDA [b]	A23, A04
HL7 CDD [c]	A25, A23
DICOM [d]	A07
HL7 V2.x and V3	A01
HL7 HQMF [e]	A23
HL7 CCR [f]	A26
HL7 VMR [g]	A03
CEM [h]	A05
IHE [i]	A27
Master data standardization and translation (MDST) model	A02

Source: Created by the author.

[a] ISO 13606 <http://www.en13606.org/information.html>
[b] Clinical Document Architecture
[c] Continuity of Care Document
[d] Digital Imaging and Communications in Medicine <https://www.dicomstandard.org/current>
[e] Health Quality Measure Format
[f] Continuity of Care Record
[g] Virtual Medical Record
[h] Clinical Element Models <http://www.clinicalelement.com>
[i] Integrating the Healthcare Enterprise <https://www.ihe.net/>

to perform in a joint work. Most of the studies had related advances toward a semantic dataset choosing standards to achieve semantic interoperability, as shown in Table 9.

Overall, the solutions proposed by the studies highlight a trend toward adopting open and international standards to meet data integration, exchange, and sharing demands. Besides exchanging data and ensuring semantic interoperability concerns as in Moreira et al. (2018), the authors had developed a framework combining ontology resources to predict high-risk situations in pregnancy. Additionally, the article contributed a summary analysis of 3 open standards, openEHR, ISO 13606, and HL7 CDA, showing their advantages and disadvantages. On the other hand, the studies Maldonado et al. (2020), Martínez-Costa, Menárguez-Tortosa and Fernández-Breis (2010), Duftschmid, Wrba and Rinner (2010) seek solutions for using openEHR and ISO13606, two open standards and similar in definition. This approach to approximation of the standards would allow the normalization of data. Although both standards use the ADL (Architecture Description Language), several differences in their types and definitions still need harmonizing.

Several works have explored applications slightly different from the known target of standards, not just EHR systems development. For instance, Yang et al. (2019)

presented a methodology to represent the dependencies among data elements, concepts, and archetypes on a three-level Bayesian network and used the inference process to discover relevant archetypes. This approach showed relevant results compared with other tools. On the other hand, the authors in Yuksel et al. (2016) aimed to improve the current post-sale drug surveillance process and ensure patient safety and reliability. The data come from voluntary reports (to wit spontaneous reports, yet just about adverse incidents). Adopting EHRs would allow for tracing a patient's medical history and predicting potential risk factors. Following a different vision, in Sinaci and Erturkmen (2013), the authors have developed a federated Metadata Registry/Repository (MDR) – a metadata database of data combining Common Data Elements (CDE) and HL7 CCD (Continuity of Care Document) models, proposing extensions to ISO 11179. Moreover, it was implemented in Marcos et al. (2015) the Health Level 7 (HL7) Virtual Medical Record (vMR) as a component service-based that aims to collect patient data from different databases to allow the use in EHR data clinical decision support as a gateway between data sources and components.

3.1.5.2 SQ2 – What are the terminologies or health repositories used?

Terminologies and vocabularies can be understood as extensive collections of terms for a knowledge domain, making the language common. Terminologies aim to prevent local expressions, neologisms, and human typing from entering the EHR, thereby adopting formal classifications, e.g., diseases, events, procedures, and specimens. Healthcare institutions share sensitive information, and systems must convey and not miss the meaning. One way is to build a local repository and manage an environment that applies proprietary concepts. However, adopting international terminologies ensures that using global terms and other classifications will have the same meaning on the other side – to any receiver. Table 10 presents the terminologies usually adopted by the studies.

While adopting standards is essential to achieve interoperability in EHR and effectively exchange information between different providers, the next level requires sharing knowledge and adding semantic value. A common language makes information comprehensible among those who sent and received it, allowing inferences in data and creating new connections from existing data. According to Alexopoulos (2020, p.3), semantic modeling essentially means linking words and terms to their senses, which is your main challenge.

Adopting a terminology engages more than one HIMSS[2] level, as it condenses structural decisions regarding the system design, team, and organizational choices. The institution's medical staff must accept adherence, use the terminologies, and publicly

[2]HIMSS <https://www.himss.org/resources/interoperability-healthcare>

Table 10: The following are the international terminologies and classifications applied by the studies aiming at a shared vocabulary focusing on keeping the real meaning.

Name	**Reference articles**
ICD-9 [a]	A23, A18
ICD-10 [b]	A23
SNOMED-CT [c]	A24, A06, A23, A11, A25, A03, A04, A20, A22, A14
LOINC [d]	A16, A03
MeSH [e]	A06
MedDRA [f]	A23
WHO-ATC [g]	A23
ChEBI [h]	A07
FMA [i]	A07, A22
BTL2 [j]	A20
BioTop	A22
Relation Ontology	A22
CDE [k]	A25

Source: Created by the author.

[a] International Classification of Diseases-version 9
[b] International Classification of Diseases-version 10
[c] SNOMED Clinical Terms
[d] Logical Observation Identifiers Names and Codes
[e] Medical Subject Headings
[f] Medical Dictionary for Regulatory Activities
[g] World Health Organization Anatomical Therapeutic Chemical
[h] Chemical Entities of Biological Interest
[i] Foundational Model of Anatomy Ontology
[j] BioTopLite2
[k] Common Data Element

reflect that choice. After compliance with the organization and team, the information conforms to a standard and is shared with other providers.

3.1.5.3 SQ3 – What are the approaches used?

The selected studies reinforce the advantages of adherence to standards for achieving semantic interoperability. Exploring new approaches involving technologies and health standards to extract better results from consolidated standards has shown growth. Likewise, some studies demonstrate the use of semantic web technologies to meet information extraction and harmonization demands. The studies had offered more than one goal, in Table 11 present the principal purpose of the selected studies.

Standards to allow semantic interoperability have been a common choice in electronic health records systems. The papers selected for this review strengthen that scenario and justify solving problems by developing solutions for this purpose. An initiative regarding making health records available for secondary use in Legaz-García et al. (2016), Yuksel et al. (2016), also reinforces a consequent advantage of implementing semantic interoperability since the improved data quality is part of the process. These works are also strongly related to the application of representations of clinical models Legaz-García et al. (2016), Lezcano, Sicilia and Rodríguez-Solano (2011), Taweel et al. (2011), OWL semantic structures, where they also serve as semantic mediators Taweel et al. (2011), Bono et al. (2011), Sinaci and Erturkmen (2013), whereas Urovi et al. (2012) also added an agent-based system to coordinate the community IHE.

Hundreds of biomedical ontologies are available in OWL format in the BioPortal repository, including many medical terminologies, which justified using ontologies in some studies to naturally convert clinical models, data, and other terminologies to this representation format. According to the authors in Legaz-García et al. (2016), exploring data semantic representation in ontologies is justified because other structures usually have explicit connections between data, and ontological systems allow reasoning to build different meanings. For this reason, representing EHR metadata in ontologies typically appears as a good choice for semantic goals.

Moreover, the use of ontologies has also proved promising for mapping scenarios of rules data access. As demonstrated in Yang et al. (2019), an agent coordination infrastructure uses an OWL (Web Ontology Language) ontology to map the access rules of organizations from the community to the data. On the other hand, the authors in Ellouze, Bouaziz and Ghorbel (2016) developed an automatic extraction from semi-structured data using ontologies. They represented those concepts into ontologies of clinical vocabularies - another successful use of ontologies.

In line with the sharing and reuse of existing clinical models, some studies propose creating automated interfaces based on archetypes from the openEHR and ISO13606

Table 11: The studies had more than one objective, so this table separates them into applications and proposed approaches to meet interoperability demands in health records, categorizing them according to the principal approach to developing the contribution.

Article contribution to solve interoperability problems	**Reference articles**
An ontology and rule-based clinical models mapping approach to solving classification/ recommendation problems in Clinical Decision Support System (CDSS)	**Rules:** A04, A14 **Ontology:** A18, A14
An ontological-based representation (a common mediator among clinical data, terminologies, and clinical models) or semantic web-based	**Intermediary ontology:** A11, A02, A19, A05, A20, A27, A23, A26 **Other semantic web tech:** A24, A06, A23, A25, A22
An ontological-based representation (clinical models and clinical data)	**Clinical models:** A06, A14, A26 **Clinical data:** A14, A22
Workflow approach as Cloud-based (component-based architecture) or Service-based framework	A01, A03, A16, A08
Bayesian network-based approach to retrieve clinical artifacts (archetypes)	A28
A collaborative framework to create clinical models by domain experts (less technical information more domain knowledge)	A15
Metadata standards-based framework to enrich artifacts with controlled terminologies and semantic labeling	A01, A07, A26, A25
XML-based EHR extract framework to represent archetype structure and constraints in XML and XML schema	A09, A17
Building automatically interfaces from archetypes by XML representations	A10, A21
IHE-based approach jointly agent-based framework – a solution to exchange patient data between a community	A27

Source: Created by the author.

Table 12: The different semantic web technologies used in studies to solve semantic problems.

Semantic web technologies	**Reference articles**
LOD [a]	A16, A06, A25
OWL [b]	A24, A18, A19, A16, A06, A11, A25, A07, A05
OWL-DL [c]	A20, A22, A14, A27
OWL-S [d]	A26
RDF [e]	A23, A16, A02, A25, A07
SKOS [f]	A23, A25
SPARQL [g]	A23, A07, A20, A2
XQuery script	A16
N3 [h]	A23
SWRL [i]	A14

Source: Created by the author.

[a]Linked Open Data
[b]Ontology Web Language
[c]Ontology Web Language-Description Logic
[d]Ontology Web Language-Services
[e]Resource Description Framework
[f]Simple Knowledge Organization System
[g]Protocol and RDF Query Language
[h]Notation 3 <https://www.w3.org/TeamSubmission/n3/>
[i]Semantic Web Rule Language

standards, such as Kalogiannis et al. (2019), Menárguez-Tortosa, Martínez-Costa and Fernández-Breis (2012). On the other hand, some studies have identified novel architectures of service involving workflows in cloud environments, aiming to enable a tool to set resource pipelines, as presented in the works Sáez et al. (2013), Legaz-García et al. (2016), Maldonado et al. (2020), Lezcano, Sicilia and Rodríguez-Solano (2011) to access health data. Follow Table 12, present the semantic web technologies used by the studies.

Despite being commonly used for semantic representation, semantic web technologies have broad applicability. Some studies have explored the representation of archetypes, rules, and relationships between different reference models (RM) of the openEHR and ISO136060 standards. The authors in Taweel et al. (2011), Marcos et al. (2015) have represented the reference model and archetype constraints into OWL ontology, which describes the instances and allows the maintenance of only one information (removing repeated instances) to keep a relationship between both RM and archetype.

Exploring other advantages Maldonado et al. (2020), Yuksel et al. (2016) presented in how RDF and OWL can be used to represent EHR data and its relationships – these transformations through the use of the Semantic Web Integration Tool (SWIT), as shown previously in Table 12. Additionally, they chose a graph database; the first used Neo4J Graph Database, and the last chose Virtuoso. However, both studies allow using Linked

Open Data (LOD).

The studies also showed a trend of combining health standards and semantic web technologies to achieve semantic interoperability. There is no consensus about adopting ontologies type using ontology resources, such as OWL and RDF ontology. Both aimed to represent the reference model and archetypes into ontologies, which allowed enriching by the ontological resource adding semantic relations. As reinforced by Demski, Garde and Hildebrand (2016), the Archetype Definition Language (ADL) usually represents archetypes and has a more syntactic orientation. Thus, it has disadvantages in achieving semantic interoperability, justifying the combining ontologies and clinical models. Also, the work presented in Marco-Ruiz et al. (2015) used ontologies combining Semantic Web Rule Language (SWRL), and these rules aimed to allow applying logical deductive reasoning between archetypes and terminologies used.

The systematic review focused on understanding the different approaches proposed to achieve semantic interoperability through health standards, and the studies showed some different choices around the database used, indicating different results when selecting a specific database. Although there are unusual and justified databases in the studies as to the reason for selection, we also kept the usual databases. Then, Table 13 present the database used by the studies.

The selected studies the not show a clear trend as to whether the health standard influences the choice of database, although it is possible to assess some characteristics. For example, the Virtuoso multi-model bank appeared in solutions using openEHR and HL7, not showing a dependency relationship with the standard type. However, a graph database allows ontologies querying if existent, reinforcing a possible hybrid solution with semantic web tools.

On the other hand, ARM and Think!EHR presented a framework designed to support archetype storage and allow querying in ADL[3]. Therefore, there is a dependency of choice when using ISO 13606 and openEHR standards. The other storage structures did not show a strong relationship or dependence on health standards, and few studies discussed the type of database adopted by their solutions.

3.1.5.4 SQ4 – What are the main security concerns used?

One of the concerns sharing data is security, such as using anonymized data and methods for de-identification. Furthermore, the exchange between different organizations must encompass safe policies, even to exchange among proprietary systems — safety issues such as maintaining security and integrity without breaking the quality of patient data.

The selected studies did not show a constant concern about data protection laws, which is a plausible consequence of the scientific research character since conducted generally in

[3]Archetype Definition Language

Table 13: The different storage tools used in the solutions developed.

Storage tool	**Description**	**Reference articles**
Virtuoso [a]	Data virtualization platform of knowledge graphs and there is an open-source version. Allows using SQL and SPARQL.	A06, A25
Neo4J[b]	Native graph database and open source, and allow to use Cypher (a graph query language).	A16
HBase [c]	It is a Hadoop database, a distributed, scalable, big datastore.	A01
ARM [d]	A proprietary approach capable of generating relational databases using archetypes and templates for archetype based EHR systems, Archetype Relational Mapping (ARM) persistence solution.	A15
Jena TBD [e]	A component to store RDF and query using SPARQL.	A25
BaseX [f]	High-performance XML database engine and a highly compliant XQuery and open source	A03
Oracle [g]	A multi-model database	A10
Think!EHR [h]	A big-data, high-performance platform designed to store, manage, query, retrieve and exchange structured electronic health record.	A17
PostgreSQL [i]	An open-source object-relational database.	A27

Source: Created by the author.

[a]Virtuoso <https://virtuoso.openlinksw.com/>
[b]Neo4J <https://neo4j.com/>
[c]HBase <https://hbase.apache.org/>
[d]Wang et al. (2015)
[e]Jena TDB <https://jena.apache.org/documentation/tdb/>
[f]BaseX <https://basex.org/>
[g]Oracle <https://www.oracle.com/index.html>
[h]Think!EHR <http://www.marand.com/thinkehr>
[i]PostgreSQL <https://www.postgresql.org/>

Table 14: The Table presents the usually followed guidelines inside academic and controlled environments to use data from the real world.

Privacy concerns	**Description**	**Reference**
Anonymization on safe zone	To use the data, the data provider first anonymized all data in a safe zone inside the health institution.	A23, A22, A12, A17
User profiles and restrict access levels	According to institution agreement, researchers use the data through restricted user profiles with different access levels.	A06, A02, A26

Source: Created by the author.

controlled scenarios in academic environments. We highlight the two definitions observed regarding privacy and security concerns. In Bahga and Madisetti (2013), HIPAA (Health Insurance Portability and Accountability Act) and HITECH, and some directives from ISO 13606 applied on Moreira et al. (2018), Maldonado et al. (2020), to store and manage health data.

On the other hand, we can consider that by keeping the security policies independent of the interoperability solutions, the authors aim to make the solutions generic in terms of local laws. Each country has its rules regarding using, sharing, and storing personal data – identification, financial or medical history. Thus, by keeping research an independent decision, studies are concerned with carrying out experiments that reflect real needs using real-world data.

Health Insurance Portability and Accountability Act (HIPAA) aims to ensure rights to citizens in the U.S., such as access to health records and request a copy of their data. As a patient, he can ask to correct something wrong with its health data, safe strategies to share data, and much more. To achieve that, they applied some methods to ensure security as authentication (SSO), authorization (OAuth), identity management, securing data at rest (using 256-bit advanced encryption standard), data in transit (HTTP with SSL), and auditing using a log data on the access. The last security layer represents the EHR stored alone because that provides data without the patient's identity. Below, we highlight some characteristics regarding privacy, which some studies have indicated as necessary measures for using real-world data in their studies.

Studies that indicated the use of real-world data demonstrated some data providers' concerns to enable the use of the information without compromising the patients. Therefore, we can observe that, regardless of the study, those who used real-world data need to apply some form of data anonymization. Table 14 present the two ways system providers usually allow data from real-world to academic researchers.

When the responsibility for data anonymization rests with the provider, the institution usually executes in a safe zone. Data only leave the health institution after being completely de-identified. On the other hand, the provider may allow researchers

Table 15: Follow we list the evaluation approaches and the types of data used in the articles.

Evaluate method	**Reference articles**
Comparative analyses (traditional data vs novel data)	A23, A12
Case study (applied inside or outside health institution environment)	**Outside** A16, A06, A25, A04, A20, A22, A12, A14, A17, A27 **Inside** A07, A03, A04
Functional evaluation	A15, A24, A11, A19, A26, A21, A05
Questionnaire (end-user feedback)	A28, A23
Questionnaire (specialist feedback)	A28, A06, A23, A02
Unitary tests	A04
Usability and compatibility across browsers	A21

Source: Created by the author.

to access the system internally. In that case, the electronic registration system must have access levels for users, not allowing unrestricted access.

3.1.5.5 SQ5 – What are the evaluation approaches used?

The evaluation of an EHR involves storing data with quality, the end-users experience (health care professional), keeping complex information from specialist physicians, and ensuring the exchange of information between healthcare providers. Therefore, a scenario where the system must adapt to the routine of health professionals and facilitate data collection but meet the demands of sharing data with quality.

In this way, approaches to evaluation aim to identify measures to validate the different levels of interoperability (foundation, structural and semantic levels) that an EHR achieved and map the challenges of systematically doing it. The studies presented other evaluation methods, and Table 11 shows the different applied approaches.

Applying questionnaires as an evaluation method allows identifying which changes – according to the user and domain expert – would be more effective in improving the final solution through constant feedback. The author in Yuksel et al. (2016) had an evaluation based on ISO/IEC 25040 (SQuaRE), System Usability Scale (SUS), and Health IT Usability Evaluation Scale (Health-ITUES). On the other hand, functional tests allow the initial assessment to establish metrics to outcomes and automated tests.

However, none of the approaches assess the extent of interoperability or validate which different interoperability levels the applications have achieved – as per the levels defined

by HIMSS [4]. Therefore, evaluating interoperability, as a result, is a challenge and an opportunity for further research, looking at more effective and unique methods.

3.1.5.6 GQ1 – What is the state of art in health standards applied in health records?

There are some challenges to be faced by the health specialists and IT professionals to achieve semantic interoperability in electronic health records. The studies showed different approaches to work around known problems and proposed new technologies and strategies to the gap area. Using the two-level (also called dual model) approach is not a global view of health standards. However, it can allow a syntax and clinical data representation common between the systems Martínez-Costa et al. (2015). Furthermore, regardless of the management system adopted, this allows of making the clinical model software-independent: the specialized health professionals model the clinical concept definition and terminologies as needed. In contrast, computer professionals are concerned with the structure, architecture, and choice of technologies to enable the use of knowledge defined by the health professional.

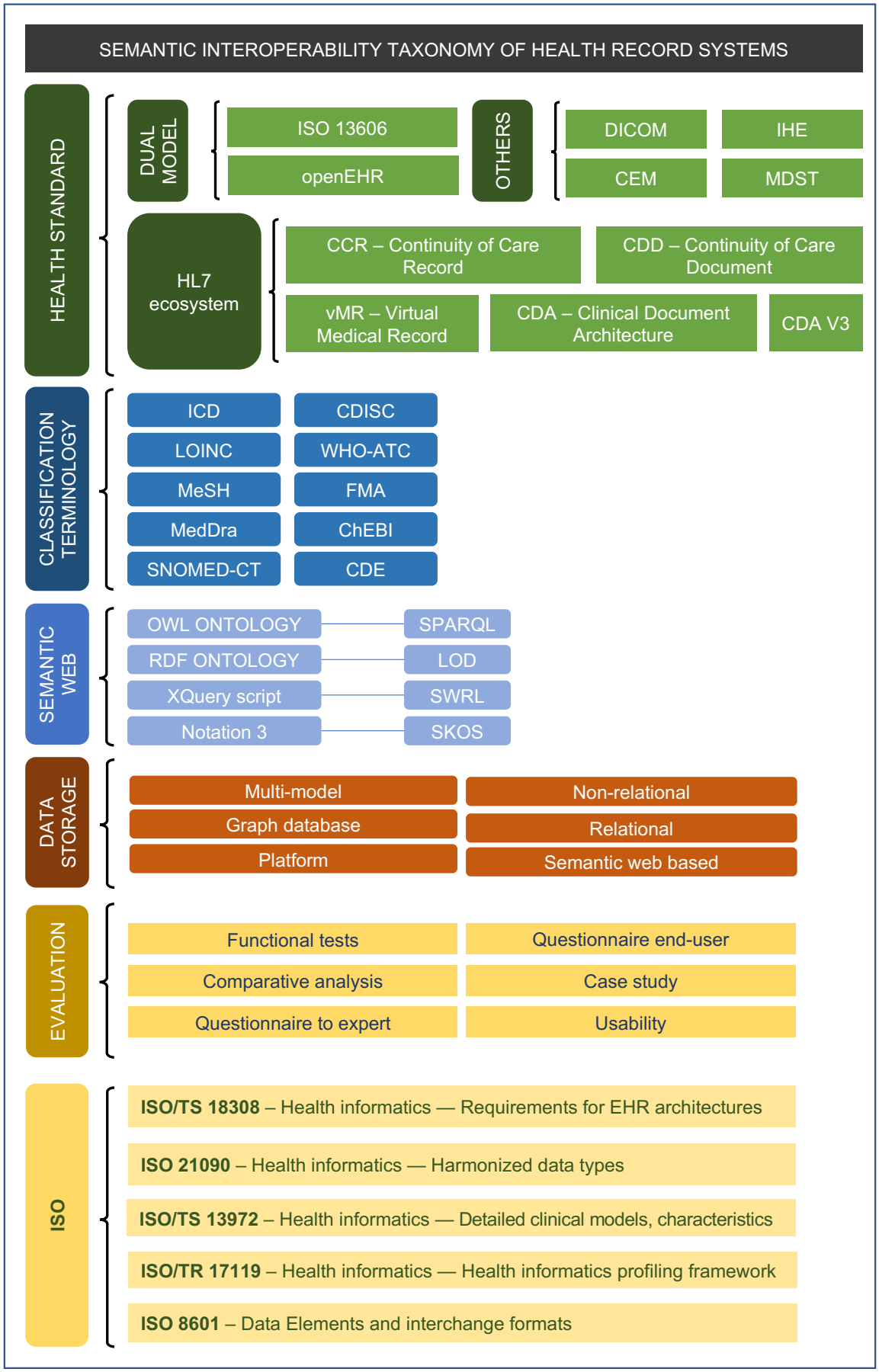
While there is not health standards consensus, we can see a trend in several studies using the openEHR. It is justified because openEHR provides a more mature public library – tools, archetypes, community support, and guides Kalogiannis et al. (2019). Another justification was about ISO 13606, which is not yet publicly available, and to develop novel archetypes to satisfied a system would take a lot of time [30]. Furthermore, to improve the adoption of the standards, many studies combined semantic web technologies and health standards. A common practice identified in the articles was the structure representation of archetypes and reference models in ontologies, combined with using rules for mapping constraints. Another use involved ontologies to represent meta-models with general information, constraint representations, and ontologies for diseases classification.

The solutions proposed by the selected articles brought different approaches to achieve semantic interoperability, Figure 7 presents a taxonomy proposal around that challenging area. Although the taxonomy is organized into five categories, the articles usually combine multiple themes to achieve the semantic interoperability challenge, so they are not exclusive sections. However, in the process of analyzing the studies, we defined six main categories: 1) Health Standards, 2) Classification and Terminologies, 3) Semantic Web, 4) ISO, 5) Data Storage, and 6) Evaluation.

A taxonomy may allow us to assess a broader scenario - e.g., what artifacts involve building an EHR semantically interoperable. We reinforce that the development of an EHR must have semantic interoperability as a goal from early project design. The decision to adopt standards and terminologies at the beginning can facilitate the development process, once the artifacts are discussed as an integral part of the project, as mandatory.

[4]HIMSS <https://www.himss.org/resources/interoperability-healthcare>

Figure 7: This figure shows a proposed taxonomy around the semantic interoperability goal in health records. The categories represent different resources exposed by the studies to solve interoperability-related problems.



The taxonomy, which encompasses accepted articles, all technologies, and standards used to address semantic interoperability and data exchange across healthcare organizations, serves a specific purpose. The selected studies have described numerous tools commonly employed to tackle interoperability issues. However, our focus is on addressing the research question posed in the article and incorporating it into the taxonomy list. It's important to note that the taxonomy is not an exhaustive map of all semantic interoperability-related tools, but rather a reflection of the tools used in the selected studies.

The first category, Health Standards, shows the standards used by the studies, with three subcategories. The dual model - openEHR and ISO 13606 inside share this architecture. Likewise, the HL7 standards have an ecosystem of standards. However, some have different goals. The HL7 organization also maintains a comprehensive list of terminologies compatible with available standards.

The choice of health standard may be related to some characteristics of the implementation team. The HL7 standard prioritizes a friendly relationship with the developer, with technical documentation and structures similar to the development ecosystem. In contrast, the openEHR proposes the opposite. Healthcare professionals have a user-friendly interface to create clinical models (in archetypes), focusing on knowledge definition. On the other hand, the artifacts defined for the openEHR standard - archetypes - have a structure that allows for in-depth clinical concept detailing, enabling interoperability at a semantic level.

Using standards inherently raises the concern of employing international terminologies to represent clinical terms and concepts. We show the vocabulary used in the Classification and Terminologies category of the taxonomy. Some terminologies adhere greatly to studies, such as SNOMED and Logical Observation Identifiers, Names, and Codes (LOINC). Most health standards allow using more than one terminology in the same clinical document with semantic binding once a healthcare organization has legacy data and treats data through many vocabularies, which is enabled in openEHR and HL7 and, consequently, in IHE.

Adopting terminologies as an inherent part of the EHR brings benefits, such as standardizing everyday expressions and ensuring semantic contextualization concerning diseases, adverse events, and general classifications. Semantic contextualization allows new connections to collected data once the structure - patient's history, lab tests, exams - has a semantic binding through international vocabularies. Unfortunately, there are still challenges when discussing the patient's history. Usually, they have an open and unstructured text field, a format more accessible to physicians and health professionals. However, an unstructured text field is not readable to the machines.

The studies use semantic tools to extract and structure the unstructured data. Although there is a low variety of semantic web technologies, as shown in the semantic

web category, the combination with health standards was almost unanimous. The most used tools often follow the definitions established by the W3C, such as OWL, RDF, SPARQL, and SKOS, as reinforced in the recent research field review Hitzler (2021). The taxonomy's Semantic Web section shows the most used ontologies, such as SKOS, OWL, OWL-DL, and RDF, as final semantic resources. We also brought to the taxonomy some standards highlighted in the studies as a reference, ISO category, for defining EHR minimum requirements, data structure, and solution architectures. The norms show the global effort towards a joint decision.

Although the studies do not explore the benefits of the choice of storage solution, some solutions are related to the type of health standard adopted. The trend towards using ontologies can impact the choice of storage solution since it is interesting to allow querying the ontological structure using SPARQL and exploring reasoning. In this scenario, Virtuoso, Neo4J, and Oracle databases present solutions that meet ontologies' storage demand.

In the Data Storage section, we map the different storage solutions to the taxonomy and categorize them according to the structure they provide. In addition, we highlight two of these solutions that specifically implement clinical document storage - archetypes such as ARM and Think!EHR. The other solutions use web-based graphical and semantic databases (W3C compliant), similar in their storage structure.

Finally, not all papers presented an evaluation format for their experiments or results obtained. However, what was striking was the rich variety of evaluation methods used. Some authors had highlighted the use of case studies in a controlled environment to apply their final experiments. Often the authors used questionnaires with end-users to evaluate the user experience. The functional tests also had a prevalence, tracking specific modules with problematic scenarios while being developed to resolve before use. In some solutions, the authors described their partial evaluation methods during the development process. They wanted to evaluate tools' accuracy or performance and not precisely the final solution, allowing for semantic interoperability. However, those methods did not consider the taxonomy because they did not influence the final discussion of the evaluated solution.

3.1.5.7 GQ2 – What are the challenges and open questions to semantic interoperability in health records?

The adoption of standards has grown as one way to ensure semantic interoperability in health record systems. However, there are many barriers to overcome in health organizations, such as legacy data, structured data, non-structured textual data, and complex internal systems that are sometimes not compatible for exchanging information. Once they overcome those challenges, institutions have shared concepts

that must preserve their meaning, even externally.

The search for interoperability at a semantic level involves combining controlled vocabularies to maintain a common meaning. For this scenario, terminologies, ontologies, and global classifications are essential. However, since there is no worldwide consensus on health standards or a vocabulary, some initiatives remain a technical issue—the absence of terminologies—when it should be organizational to define the mandated use.

On the other hand, vocabulary harmonization within EHR systems represents a promising alternative toward a feasible decision. It would allow the adoption of many controlled vocabularies and harmonize them as unique meanings. The lack of common terminologies and the extensive use of proprietary concepts inside EHR make interoperability difficult, requiring normalization processes. In other words, an effective EHR must be concerned with standardizing syntax, structure, and semantics (concept sharing). The papers presented showed concerns about the difficulty in harmonization and concept normalizing in legacy systems, as previously mentioned in Yuksel et al. (2016), Sinaci and Erturkmen (2013). Both works expose complex models and legacy data as reuse and secondary use boundaries—moreover, proposals to explore clinical research and decision support systems in Martínez-Costa et al. (2015), Sáez et al. (2013).

These foreground scenarios that future studies can explore, for instance, open health standards and international terminology knowledge, are still far from being extensively used by healthcare professionals and are only widespread in academia (educational institutions) and large software providers. Therefore, the opportunity arises to contribute with measures that mitigate the recurring difficulties of standardization, making knowledge about available resources accessible, easy to implement, and providing adequate materials to facilitate the study.

Since knowledge about standards is not widespread, few systems implement controlled terminologies, ontologies, and classifications, as they are separate from the reality of healthcare in their designs. Therefore, working on initiatives to gradually adopt controlled vocabulary daily is crucial. That lack means opportunities to propose methodologies to adopt a health standard, map possible difficulties and how to face them, and offer feasible directives for the reality of health institutions.

Interoperability as an integration focus and information exchange is part of the scope of software providers, a commonly raised concern. However, healthcare systems consider semantic interoperability a plus when it should be a prerequisite since it directly implies the quality of the stored data, which will later be used in secondary studies. Therefore, encouraging the construction of systems aiming to achieve semantic interoperability would allow exploring secondary use of EHR, data reuse, and clinical models and making clinical research within institutions a reality.

3.1.6 Literature review findings

When defining interoperability as the foremost goal, it is essential to identify the paths to achieving interoperability present challenges at different levels—technical, structural, semantic, and organization (HIMSS, 2021). This scenario begins with how the patient's information is collected by the other agents of the health institution, such as nurses, clinicians, and caregivers. Moreover, these are usually inputted in the free-text form of the EHR system without controlled vocabulary.

Considering the sensitive and high level of quality needed for health data, the biggest challenge to healthcare providers and EHR system providers is to perform and collaborate to promote initiatives to adopt structuring data by the time professionals collect and input it into the system. The choice to provide a system concerned about interoperability by the record is an Organizational decision based on the interoperability goal. However, its benefits include financial security and patient care continuity by allowing secondary use, mitigating the lack of information between care continuity, and providing a more confident scenario for professionals' shift transition.

Throughout the literature review, we identified main opportunities in the health domain related to data interoperability, besides understanding the different tools usually used to achieve interoperability at a semantic level. The results showed a close relationship between adopting healthcare standards—mainly open standards, with a two-tier approach (e.g., openEHR and ISO13606)—and supporting semantic web technologies. This approximation of semantic web technologies and healthcare standards allows approaches to semantic data representation (EHR), clinical information models (openEHR archetypes), and rules to enable optimized searches.

This strong relationship is also evident in harmonizing standards, the combination of semantic web tools for representing clinical information structures in ontologies as temporary structures to later convert the data to the final standard in a formal format such as HL7 ecosystem, OpenEHR, ISO13606, and others. This approach allows for maintaining a semantic structure of the information models. At the same time, the standard definition formats tend to have more syntactic orientations (e.g., Archetype Definition Language, language to represent the openEHR clinical information model–archetypes), which can lack data context representation depending on the standard choice.

Health standards also have a strong relationship with the guiding aspect of publishing and following ISO specifications. We emphasize mainly the data types (ISO 8601–Data Elements and) and minimum requirements for an EHR (ISO/TS 18308–Health informatics), present in several studies selected as a starting point for interoperability decisions, both highlighted on the Taxonomy proposed in Figure 7.

The taxonomy contributed to our literature review and allowed us to evaluate the leading technologies in developing a semantically interoperable system for data from

medical records. The taxonomy showed the adoption of standards as a path to interoperability and presented the different controlled vocabularies adopted that are responsible for allowing interoperability at a semantic level. In addition, the taxonomy presented the databases usually adopted by the selected works, which have a particular relationship with the type of health standard chosen. On the other hand, some studies have highlighted the difficulty concerning legacy data stored in unstructured formats and heterogeneous bases.

Therefore, this scenario presents itself as a research opportunity since digital transformation and the growing adoption of EHR systems in institutions have led to the large-scale production of unstructured data. Recent studies have explored the use of information extraction techniques such as Named Entity Recognition, Relation Extraction, and Entity Linking to structure and allow the secondary use of unstructured data– the main format collected and maintained by electronic health records systems today (MARTIN-SANCHEZ; VERSPOOR, 2014; SIVARETHINAMOHAN; SUJATHA; BISWAS, 2021).

On the other hand, according to recent works discussed, some approaches combining machine learning, Deep Learning, and NLP techniques have shown promising results in tasks for extracting information (LI et al., 2021; NEGRO-CALDUCH et al., 2021; LI et al., 2020a). Moreover, it proposed tools to extract quality and semantic information using controlled vocabularies. The studies also demonstrated a trend towards the adoption of deep learning architectures, with better results when the objective is to apply information extraction tasks, such as RNN (i.e., LSTM, BiLSTM, GRU) and CNN (HOUSSEIN; MOHAMED; ALI, 2021) besides the BERT architecture, which has information extraction in all NER, RE, EL have demonstrated a general better performance on those tasks.

Finally, the studies evaluated in the literature review reinforced the pressing need for Electronic Health Record (EHR) systems to prioritize semantic interoperability when inputted into the system. Additionally, they emphasized the challenges in improving this information collection. Transforming unstructured data into high-quality, structured data presents significant complexity for healthcare providers and EHR systems providers.

3.2 Machine learning and deep learning scenario

The growing adoption of electronic health records is a key factor contributing to the large volume of data generated for healthcare providers, including clinical histories, imaging exams, vital signs, lab results, patient evolution, and clinical notes. The digital transformation from paper to electronic records has created a path of no return, introducing a new, integrated service way for citizen/patient care.

Using EHR within institutions has brought new perspectives for using the collected data, looking forward to secondary use, clinical research, clinical decision support

systems, and other uses. Although a digital version of the EHR has most of the data non-structured or not adequately structured to enable novel research (SIVARETHINAMOHAN; SUJATHA; BISWAS, 2021; LI et al., 2021; XIAO; CHOI; SUN, 2018), that digitalization of health data brings opportunities to explore method and techniques to qualify information.

Another point to consider from the systematic review foregrounded scenarios of fragility, exposing the difficulties of an interoperable electronic health record in the health institutions' reality. The existing legacy data, often unstructured, do not have any health standard or controlled vocabulary. Moreover, since there is no integrated system, the patient's history is fragmented, with duplicate entries, which worsens at the level of system providers, who hold proprietary storage structures. It became impossible to share data among health institutions without spending an extended period for adjustments.

The review showed alternatives using semantic web technologies combined with health standards to meet the demand for semantic interoperability. For instance, the ontologies prevalence to represent reference models (RM) and clinical information models (CIM) of the health standard adopted, as shown in Table 11 reinforce the ontology data representation can improve the approach in the semantic aim (BLACKMAN-LEES, 2018; YANG et al., 2019; YUKSEL et al., 2016). Moreover, we highlight the preference for adopting semantic web technologies recommended by the World Wide Web Consortium (W3C), such as RDF, OWL, SPARQL, and SKOS, as shown in Table 12.

Furthermore, studies that explored the two-level standards showed OWL ontology's predominant use, while studies that explored available HL7 standards showed RDF ontology's primary use. Despite this, in both scenarios (HL7 and two levels - openEHR and ISO13606), the studies advocate adopting ontologies as a suitable format to explore the use of metadata and semantic relationships between controlled vocabularies. The studies regarded the possibilities of inferences through the use of ontologies, reasoning, and possibly applying rules as the main reason. Also, it is essential to obtain semantic orientation since other formats have a syntactic orientation.

The situation within healthcare organizations is still far from the controlled environment, often in research projects. For this reason, it is necessary to propose feasible alternatives to health institutions as approached by HIMSS (2021). Usually, institutions partially reach the first two interoperability layers - structural and fundamental. That scenario is reinforced by Hayman et al. (2019), Lee, Kim and Lee (2021), regarding the three levels of interoperability to improve workflows and exchange data across health systems.

Although these initiatives are from adopting previously discussed health standards, the reality is that healthcare providers present collected data in multiple formats, such as

tabulated and structured, textual non-structured, and free-form, resulting in fragmented contexts. Also, add the lack of controlled terms and any concepts definitions machine-readable–acronyms and colloquial phrases. An environment for standard adoption needs initially structured data, controlled vocabularies, and normalized concepts. Moreover, finally, select a standard that meets the institutions' interests, preferably one that allows data representation at a semantic level.

In the health domain reality, and mainly when referring to EHR systems, two main choices for data representation stand out, known as structured and unstructured health data (SIVARETHINAMOHAN; SUJATHA; BISWAS, 2021). On the one hand, structured data consists of heterogeneous representations of data in different databases, such as diagnoses, medications, and laboratory values in fixed numerical or categorical fields (LI et al., 2021). On the other hand, unstructured data refers to free-form text, non-structured, written by healthcare professionals, clinicians, nurses, and caregivers, making secondary use a challenging issue. Nevertheless, these different kinds of information are one of the reasons the unstructured textual data provide opportunities to explore different insights from the data ('NASEEM et al., 2021; KORMILITZIN et al., 2021; LAPARRA et al., 2021).

The large volume of health data has been a challenge over the years, a situation discussed by the authors Martin-Sanchez and Verspoor (2014), who highlighted concerns about the large-scale growth of health data in heterogeneous repositories non-structured. The authors also reinforce the need to propose a data management, data linkage, and data integration strategy for the existing information system, since 80% of the EHR systems data are unstructured (i.e., clinical free-text form, imaging, exams). In this scenario, the health domain has a reality that is heading towards the adoption of Big Data technologies, with processing, storing, and analyzing large-scale data needs(NASEEM et al., 2020). Furthermore, as noted by the authors Negro-Calduch et al. (2021), the growing volume of healthcare data has been the hallmark of this digital transformation in the health domain.

On the one hand, information extraction in the health domain presents several challenges because of the massive volume and data quality. On the other hand, this scenario has aroused growing interest, promoting joint research between machine learning techniques and NLP tasks focused on information extraction. Recent works to solve non-structured problems have proposed to employ NLP and ML techniques to make data machine-readable (LI et al., 2021; HOUSSEIN; MOHAMED; ALI, 2021).

In line with this vision, the literature review conducted by the authors Negro-Calduch et al. (2021), covered the period from 2010 to 2020, observing the leading technologies used to solve problems among the Electronic health record (EHR) and personal health record (PHR) systems domain. The review outcomes highlighted studies exploring information extraction and NLP techniques, followed by blockchain technologies in healthcare (ROEHRS et al., 2018). For the results, we focused on the

details mapped as main concerns to be addressed by NLP tools in the health domain: to enable secondary use of EHRs to unlock the knowledge from unstructured data in EHRs. As a result, the demand presented, the study pointed out as a usual solution, they develop deep-learning-based NLP for EHR data mining, as well as the development of ontologies.

Given the scenario of large volumes of unstructured data, Nasar, Jaffry and Malik (2021) reviewed the literature to provide an overview of the main approaches applied to extract information from textual data and structure it. In addition to discussing state-of-the-art studies, the authors divided the paper into sub-domains of information extraction: Named Entity Recognition (also known as Named Entity Extraction) and Relation Extraction. The review highlights trends in related works employing information extraction techniques. The study mapped areas where combined approaches were adopted alongside traditional NLP tasks, such as NER and RE. According to the study, there is a clear trend towards using Deep Learning approaches in conjunction with NER and RE tasks. Additionally, Transfer Learning (TL) approaches are particularly prominent in studies focused on solving NER problems, although fewer studies have applied TL to RE problems. The authors also emphasize a strong trend towards adopting supervised approaches for NER tasks.

Furthermore, the review conducted by Houssein, Mohamed and Ali (2021) emphasizes the growing prevalence of unstructured data due to the increasing adoption of EHR systems, highlighting Natural Language Processing (NLP) techniques as a critical strategy for extracting information from clinical narratives. This extraction is crucial for the secondary use of EHRs in clinical research. The review also sheds light on the current landscape of Machine Learning (ML) and Deep Learning (DL) techniques in biomedical NLP tasks, underscoring the importance of deep learning, machine learning, and rule-based methods. Their research spans studies from 2009 to 2021, identifying deep learning approaches like Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM), BiLSTM, and Gated Recurrent Units (GRU), along with Convolutional Neural Networks (CNNs) for simple concept extraction tasks such as Named Entity Recognition (NER). Similarly, (LI et al., 2021) reinforces this trend, pointing out fundamental studies that combine NLP and DL techniques, employing these architectures specifically for NER tasks in biomedical data.

Additionally, Giorgi and Bader (2020) explore enhancements to the BiLSTM architecture by integrating it with a Conditional Random Field (CRF) layer to improve performance on BioNER tasks. Their approach involves modifications such as variational dropout, transfer learning, and multi-task learning, further advancing the ability to generalize within biomedical data extraction tasks. This work demonstrates the critical role that advanced deep learning architectures play in improving information extraction from clinical narratives, a vital step toward enabling semantic interoperability

in healthcare systems.

Moreover, the authors Rivera-Zavala and Martínez (2021) proposed a comparative study to explore Named Entity Recognition (NER) for information and knowledge acquisition in pharmaceutical, chemical, and other biomedical entities clinical texts(in the Spanish language). Their first experiment used the BiLSTM-CRF model, and the second used a BERT-based architecture with promising outcomes for the NER task. However, the authors discuss the low indexing of concepts due to the significant spelling errors and abbreviations. Following the trend previously discussed around adopting RNN architectures to information extractions–Named Entity Recognition tasks in Houssein, Mohamed and Ali (2021), Li et al. (2021).

On the other hand, some studies explore the recent transformer architecture proposed by Devlin et al. (2018) for NER tasks. The authors in Hakala and Pyysalo (2019) used the BERT-Base-Multilingual-Cased[5], then fine-tune to the NER task, replacing the last layers with a CRF layer. The study carried out by Naseem et al. (2020) focuses on tasks for extracting information from the biomedical domain, adopting a hybrid approach. First, the BioBERT model to a text representation, and then using attention Bi-directional Long Short Term Memory (BiLSTM) with Conditional random field (CRF) for the BioNER task. Furthermore, the BioBERT-MRC, proposed by Sun et al. (2021), adopted BioBERT to perform BioNER in the Machine reading comprehension (MRC) framework. Formally, the MRC task focuses more specifically on classification tasks, but some studies have explored its application in NER tasks with positive outcomes.

Considering this scenario, we identified a research opportunity to propose and develop methods for semantic interoperability, particularly focusing on integrating Natural Language Processing (NLP) techniques and ontology-based frameworks. This approach would facilitate the automatic structuring of clinical unstructured data, enabling more efficient information retrieval, enhancing data accuracy, and supporting seamless integration across healthcare systems. This proposal aligns with the broader need for health informatics to advance toward more intelligent, context-aware systems that can manage unstructured data more precisely, improving patient care and data sharing across platforms.

Finally, we highlight some initiatives that aim to promote information extraction tasks on Portuguese language domains by the authors Lopes, Teixeira and Oliveira (2019), Schneider et al. (2020), Souza, Nogueira and Lotufo (2020), Oliveira et al. (2020). As reinforced by Kormilitzin et al. (2021), initiatives for annotating data in other languages–non-English languages–have grown in recent years due to the need for corpora annotated on the target language to improve outcomes in Neural NLP models.

In this sense, this research proposes a model to promote semantic interoperability in non-structured health data. We propose acquiring non-structured data from healthcare

[5]BERT-**Cased** type because it preserves the case and accents of the character.

institutions and applying data processing (including ML, DL, and NLP techniques) to represent and structure data into a semantic and standardized format. Therefore, the next chapter presents the proposed model and describes the components needed to achieve a semantic interoperability goal in electronic health records.

4 PROPOSED MODEL

This chapter describes the proposed model and the components required to build a semantic model for health data interoperability, considering the solutions proposed in recent literature and related works to address known challenges. Furthermore, this research is related to the MyDigitalHealth project[1], and we direct our efforts according to the COVID-19 case study delineated from the databases made available by the healthcare provider partners.

We designed the model by focusing on two main scenarios. First, we conducted a literature review on semantic interoperability, identifying research gaps based on the needs of healthcare providers. Additionally, our hospital partners actively participated in meetings to discuss their necessities, challenges, the lack of data standards, processes for collecting and storing data during the pandemic, and the reality of unstructured data as an obstacle to integration and interoperability. The proposed model aims to promote data interoperability among health institutions. Therefore, it focuses on hospital EHR realities, where the collected data are presented in structured and unstructured formats.

In the proposed model, we sought to design an interface aligned with the approach of handling interoperability in layers, considering aspects to be addressed at the most appropriate times without anticipating unnecessary processing or difficulties.

The first layer concerns infrastructure, network protocols, and the technical environment that composes the technological ecosystem. The second layer concerns the format of the data being exchanged. In contrast, the third layer concerns the choice of the appropriate protocol to represent data in the selected format, which is handled in the syntactic and semantic layers through a healthcare standard. This modular approach reflects best practices recommended in various areas of healthcare interoperability. On the other hand, all decisions about data governance, legal law about sharing personal data, levels of access, and permissions are made at the fourth layer.

The ISO/IEC 21838 standard on top-level ontologies (TLO) exemplifies how semantic interoperability can be addressed modularly, dealing with different aspects of meaning in separate layers. Similarly, FHIR (Fast Healthcare Interoperability Resources), developed by HL7, adopts a layered structure for interoperability, defining resources that address specific issues independently but in an interconnected way, allowing for more seamless integration across different healthcare systems.

Moreover, the OpenEHR model promotes the separation of concerns through different levels of abstraction, such as archetypes and templates, which facilitate subsequent processing and semantic interoperability, as discussed by Beale, Heard et al. (2007). These practices are also reflected in the IEEE 11073 standards for personal health devices, which use a layered approach to ensure effective communication and data

[1] In Portuguese MinhaSaudeDigital (MSD).

exchange between devices and healthcare systems.

Through the proposed model, we perform information extraction tasks applied at different levels, aligned to interoperability related to layers: technical, structural, and semantic. As a result of these layer's integration, at the end of those layers, the model aims to deliver structured data with semantic value in the Organizational layer. Consequently, the model has some processing data stages to achieve interoperability at the semantic level—i.e., structuring and contextualizing health data and aligning with terminologies and standard choices.

The model employs Named Entity Recognition (NER) techniques in non-structured data to extract key terms as entities and ensure shareable data in a formal structure. This research provides a valuable complementary deliverable, a methodology to annotate small datasets.

4.1 Overview about data challenges

This research is part of the MSD project, which involves seven hospital partners in Rio Grande do Sul, Brazil, to conduct the study and explore real-world health data. In this context, we emphasize the participation of healthcare professionals as agents of tacit knowledge within these institutions. All suggestions and needs regarding data collection, management, and usage within the institutions, as well as their challenges with data sharing, guided us in designing a solution proposal aligned with real needs. Additionally, this relationship provided direct access to raw health data, reflecting the everyday reality, and offered the opportunity to propose measures to address critical situations more effectively.

This scenario helped us understand the distinct sectors involved in the attendance patient process (the study case as COVID-19 data). Hence, we highlight the importance of applying studies to data from real-world, with raw format from the institutions, not to mention the possibility of obtaining data from Brazilian Portuguese, which enable understanding challenges regarding the demand for the Portuguese language tools. To that end, we organize meetings to discuss intern demands with the partners, which are: Grupo Hospitalar Conceição[2], Santa Casa de Misericórdia[3], Hospital Ernesto Dornelles[4], Hospital de Clínicas[5], Hospital Moinhos de Vento[6], Hospital Universitário de Santa Maria[7], and Unimed Central de Serviços RS[8]. As a result of this approximation with

[2]Grupo Hospitalar Conceição <https://www.ghc.com.br/>

[3]Santa Casa de Misericórdia <https://santacasa.org.br/>

[4]Hospital Ernesto Dornelles <https://www.hed.com.br/>

[5]Hospital de Clínicas <https://www.hcpa.edu.br/>

[6]Hospital Moinhos de Vento <https://www.hospitalmoinhos.org.br/>

[7]Hospital Universitário de Santa Maria <https://www.gov.br/ebserh/pt-br/hospitais-universitarios/regiao-sul/husm-ufsm>

[8]Unimed Central de Serviços RS <https://www.unimed.coop.br/web/centralrs>

hospital reality, we identified the following critical scenario:

- Data heterogeneity includes multiple internal systems from different providers and data collected by physicians, nurses, and attendants in each system also discussed by the authors in Maldonado et al. (2020), Kalogiannis et al. (2019).
- The growing adoption of EHR systems, which lead to the open textual entries and unstructured fields when there is not adequate fields to describe patient status and generate duplicate entries.
- Legacy data in proprietary or unstructured formats aligned the absence of terminologies in the EHR, reinforced by the studies Marco-Ruiz et al. (2015), Duftschmid, Chaloupka and Rinner (2013).
- Difficulties in sharing data because of their unstructured reality and the knowledge fragmentation of in heterogeneous data sources (LI et al., 2021).
- Regarding the LGPD initiative and other data policies to keep patient data secure, we highlight the difficulty of conducting studies on healthcare data from the real world, which requires efforts to anonymize and replace before.

In this perspective of joint work to observe characteristics and needs from the real world, we can highlight some works that corroborate the decision of this partnership, with promising results over the years. The studies used several kind of data, with different sources, coming from the real-world, such as (LOZANO-RUBÍ et al., 2016) pregnancy data, (YUKSEL et al., 2016; MALDONADO et al., 2020) colorectal cancer screening protocols and lab tests, (SINACI; ERTURKMEN, 2013) pharmacoepidemiology data, (MARCOS et al., 2015) infants affected by Cerebral Palsy, (MARTÍNEZ-COSTA; MENÁRGUEZ-TORTOSA; FERNÁNDEZ-BREIS, 2010) coronary computed tomography angiography, (LEGAZ-GARCÍA et al., 2016) atrial fibrillation (AF), (BONO et al., 2011) cite mellitus, (DUFTSCHMID; WRBA; RINNER, 2010) chronic heart failure, (MARCO-RUIZ et al., 2015) abnormal reactions and allergies patient's data. Also, other studies used only results from lab tests as (LEGAZ-GARCÍA et al., 2015) radiology data, (MIN et al., 2018) histopathology reports, (MARTÍNEZ-COSTA et al., 2015) test results of pertussis and salmonella.

Considering this relationship with the health partners, we consulted about the reality of their data, the partners involved in the research informed us that it is structured and non-structured. There are open textual fields in the health record system, and they fill in the patient's information monitored during the shift. In this way, professionals make manual entries in a report or direct specification of drugs administered. Besides, when the inpatient needs a specialist appointment, the specialists make their manual entry into the system - which means something like their specialty, an overview of the patient's status,

the reason for the appointment, followed by a prescription report of what should be given or performed, such as prescribing medications or procedures.

Furthermore, analyzing the medical notes recorded, we identified duplicate entries by patient. That situation can happen when the responsible professional needs to update patient information. The professional duplicates the previous record so as not to modify it and updates it with the new information below. For example, sometimes the entries appear in any order and with different content, such as (1) Symptoms, date since the first symptoms, the prescription and procedures; (2) Shift, patient name, type of call, patient status, symptoms, prescription, and procedures; (3) Type of specialist, patient status, date since the first symptoms, procedures performed, medication administration, procedures to be performed. Likewise, these entries, possibly made by different people in charge, usually cause semantic inconsistency since professionals have additional terms for similar events but different specialties.

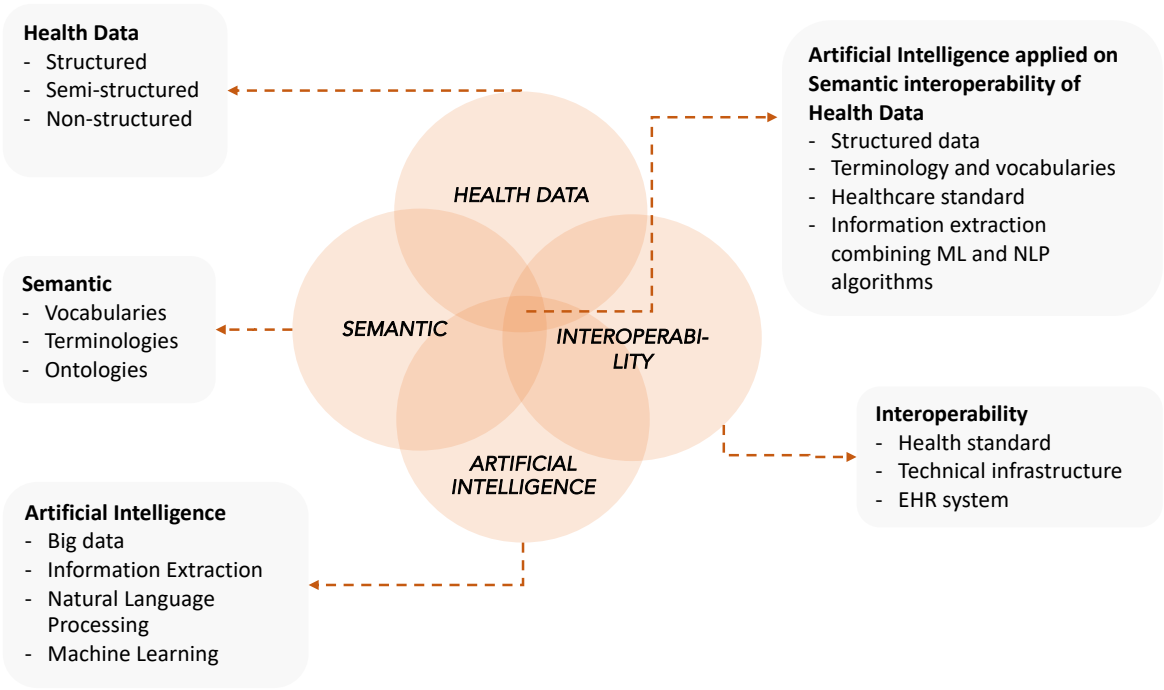
Therefore, this scenario makes the information unreadable to machines. Concerning patient data, we may exclude some data that may be presented structured, such as patient identification, name, age, birth date, profession, and marital status. However, this information is useless for sharing knowledge regarding the patient's general condition, a situation that we are interested in sharing among healthcare institutions.

4.2 Architecture of the Proposed Model

Over the years, research on healthcare data, interoperability, and semantics has expanded, and both unstructured and structured health data have been the focus of various studies Malm-Nicolaisen et al. (2019), Laar et al. (2020), Palojoki et al. (2024) aimed at understanding the EHR landscape and proposing approaches to optimize its structure, enabling clinical research, secondary use Negro-Calduch et al. (2021), and information exchange among institutions. Figure 8 illustrates the intersection of this research with the goal of achieving semantic interoperability, highlighting the challenges across the healthcare domain and our objective within that intersection—applying Artificial Intelligence to the semantic interoperability of healthcare data.

To propose the model, we analyzed the entire scenario presented by the partner hospitals. We had unstructured textual data, no pre-established storage rules between institutions for adopting interoperability standards, and no formal definition for using terminologies or controlled vocabularies. At that point, we began identifying the obstacles, focusing on the characteristics of unstructured textual data and the quality of information, including clinical contexts such as laboratory results, vital sign measurements, exams, medications, procedures, and other relevant medical information in the records. Our analysis revealed critical scenarios that presented significant obstacles to achieving semantic data interoperability, and we highlighted several

Figure 8: The intersection of the research challenge.



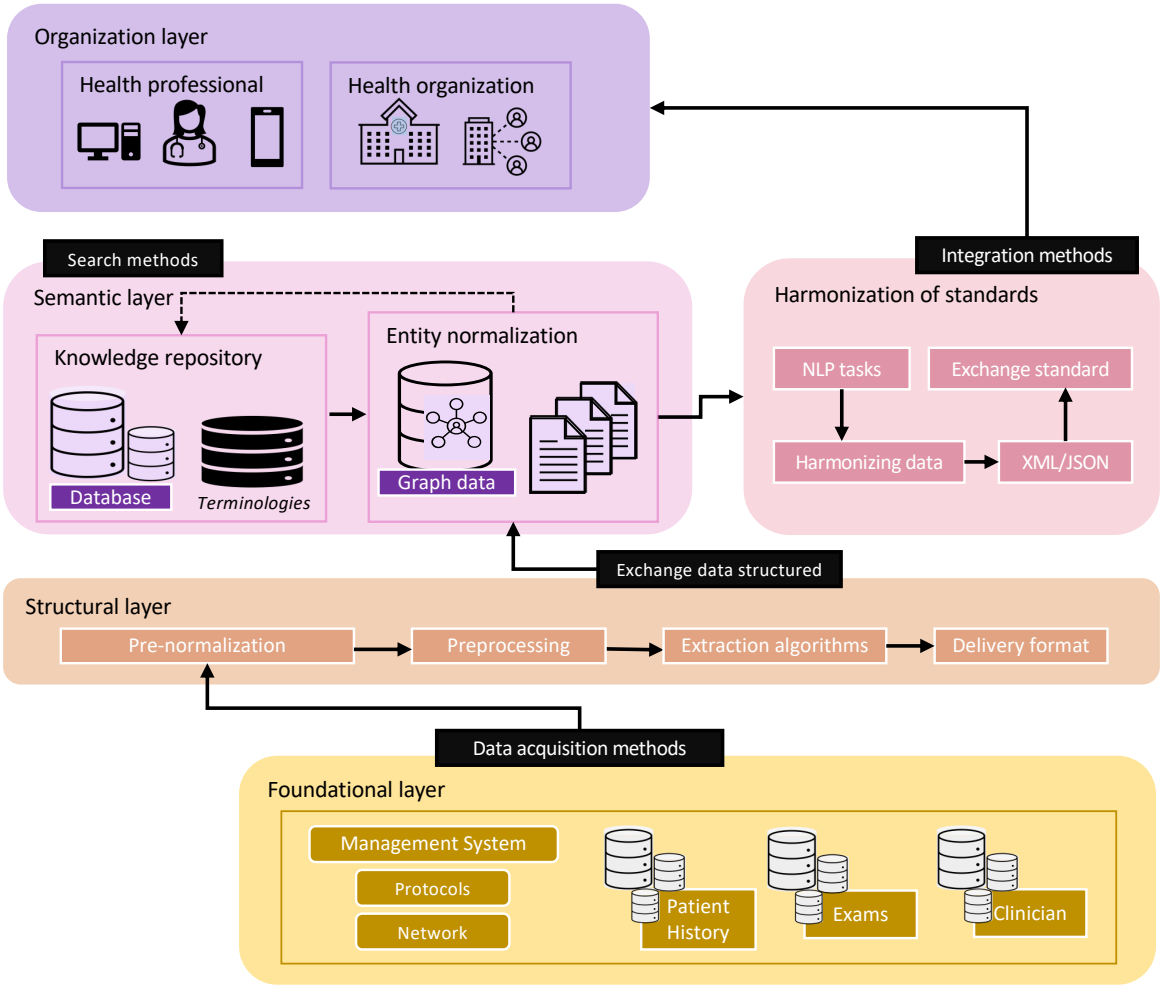
Source: Created by the author.

technical impediments related to the partners: (1) the methods hospitals used to retrieve and format stored data, (2) the quality of the information retrieved for sharing, and (3) the mechanisms in place to enable the sharing of information among professionals and institutions. Since data interoperability is addressed in layers, it was crucial to analyze and understand the specific nature of these issues to propose appropriate techniques for addressing them at each layer.

In this scenario, we adopted a modular approach to address the problems at different layers of the interoperability model, ensuring that each issue is tackled within the appropriate layer. Following the recommendations of HIMSS (2021), addressing the various aspects of interoperability from a layered perspective can facilitate a smoother transition toward adopting interoperability standards. Figure 9 presents the conceptual model proposed for this research. We identified the difficulties and gaps in the data from the partner hospitals and outlined the latent need to interoperate the data provided to this research among the hospitals.

Given the data and the challenges that emerged, this research concentrated on exploring how to structure unstructured textual healthcare data, which are a rich source of patient information and remain one of the most commonly used formats in electronic health records. Healthcare professionals preferred this format due to the ease and speed of using natural language to input data. However, these systems, which stored unstructured information, often lacked sophisticated or effective methods for data search

Figure 9: The proposed model to achieve semantic-level data interoperability in non-structured data.



Source: Created by the author.

or retrieval in subsequent encounters, hindering standardization, sharing, and ultimately the continuity of care. To address this, we experimented with techniques that extracted information from health records and provided a standardized structure using international healthcare standards to ensure semantic interoperability among hospitals.

On the other hand, non-structured data needs special attention to maintain the quality of information once it originates from manual entries. The use of acronyms, negation adverbs, and grammatical errors, combined with cultural differences and variations in writing styles, leads to significant discrepancies in the data, alongside the potential for human error during data entry or typing as mentioned by Rocha et al. (2022). Consequently, we related these scenarios to the institutions' routines, seeking solutions closer to their realities. We proposed an interoperability model to meet these fundamental difficulties in working with problems from within institutions. In addition, we brought up the possibilities that a semantic interoperability model can bring inside institutions.

In our proposed model, we used NLP techniques and extracted information by applying Named Entity Recognition (NER). The most common NER tasks involve extracting general entities like events, names, addresses, people, and places. To that end, we experimented with techniques to extract contextualized information from unstructured data by applying BERT (Bidirectional Encoder Representations from Transformers) in downstream NER tasks within the healthcare domain (as previously described in Section 2.5).

The proposed model consists of four layers, with the semantic layer extending across nearly five. The first layer, the Foundational layer—also referred to as the Technical layer, represents the healthcare provider's environment, acknowledging that we cannot access their internal ecosystems due to security and privacy concerns. Therefore, the model proposes an intermediary method to facilitate data acquisition from this environment at the boundary. This method does not require a specific format; rather, it serves as an interface to securely export unstructured data, which was standardized by the semantic model.

The second, the Structural layer, represents a set of processing data phases since the acquisition method. The model has non-structured data to promote some structure to data through these phases. That process starts with the layer receiving data from the data acquisition method and later applying pre-normalizing, preprocessing to information extraction, and disambiguation techniques. Aiming to extract entities and relations, avoid ambiguous concepts, and deliver data to the next layer.

The Semantic layer data has a structure and extracted relations with contextualized entities. However, usually in the medical domain needs to adopt a controlled vocabulary and terminology. This layer employs entity normalization techniques to map entities extracted to a unique term. Also, the semantic layer has an extended module, where

the model proposed standard harmonization, aiming to hold structured data into an international health standard. That module focuses on the interoperability capacity to make sharing between health institutions.

The last organizational layer represents the data governance aspects of the model, and it applies laws, regulations, and national definitions to the data extracted before it is shared externally as an outcome of the model. On the other hand, this layer finishes the activities, consolidating the process of structuring and qualifying data to allow interoperability. On this layer, managers and coordinators adopt the new processes into the institution and the entire team to exchange, share and interoperate data across healthcare providers.

4.2.1 Foundational layer

Also called the Technical layer covers the applications and infrastructures to relate systems and services, covering data transport aspects, communication protocols, and the network. It represents the infrastructure behind the first contact out of health institution context. As defined by Commission et al. (2017), technical interoperability should be ensured, whenever possible, via the use of formal technical specifications and recommend the use of open specifications.

The Foundational layer in the proposed model encloses two key aspects. On the one hand, interoperating data involves defining the use of technologies and organizing the technological infrastructure, which is decisions exclusively under the healthcare providers' own. On the other hand, interoperability must be ensured regardless of the technological choices made by healthcare providers, whether those choices are driven by cost-benefit analysis, partnerships, technical team expertise, or other factors.

In other words, technical interoperability involves network protocols, ensuring that systems are connected and can communicate, regardless of the hardware or operational systems used. As well as the use of Application Programming Interfaces (APIs) as routines, protocols, and tools that allow different software systems to communicate. In the healthcare context, standardized APIs enable systems such as electronic health records (EHRs), mobile health applications, and clinical data management systems to exchange information. The technical layer supports message exchange using standards like HL7 V2 (Health Level 7 Version 2), which defines the structure for exchanging healthcare information. Another example is DICOM (Digital Imaging and Communications in Medicine), used for the exchange of medical images.

Even though the two standards mentioned above, DICOM and HL7 V2, are healthcare standards, they address the needs of the technical layer, specifically acting as intermediaries between the technical layer and the higher layers, but focusing on how data is structured and organized for exchange. For example, while HL7 V2 focuses more

on the mechanics of data exchange (technical layer), HL7 FHIR emphasizes the form and structure of the data to guarantee meaning, ensuring that the context from the sender is the same as the receiver, and understood uniformly by different systems (structural layer). Also, in this layer, the security protocols are defined, such as the use of TLS/SSL (Transport Layer Security/Secure Sockets Layer), as well as considerations for authentication, auditability, and data authorization.

In this layer, our focus on having an intermediate component, an interface, to ensure to communicate data. Therefore, in the Foundational layer's first iteration, we proposed Intermediate Data Acquisition component.

4.2.2 Structural layer

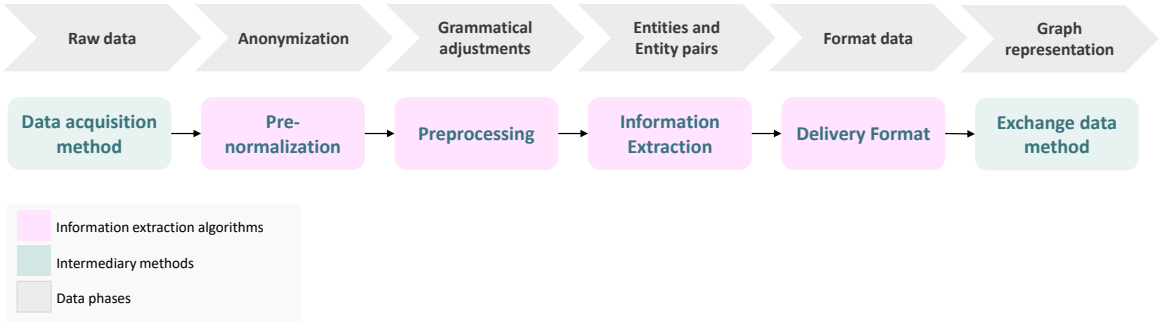
This layer has the first interaction with data; however, we are still not talking about structuring it. At this first moment, it is necessary to provide an intermediate component to receive the data, which we can receive in different formats, some often used, such as XML, JSON, TXT, and CSV. Since each health institution presents different realities for information sharing, it becomes impracticable to request a standardization effort, which may cause delays in receiving data or deciding not to participate from partners. That is why this intermediate method is at the structural layer boundary, as it has the versatility to satisfy many formats. With the received data, we evaluated the format received and treated them. Then, we processed it according to the foreseen phases.

This layer defined the components identified as fundamental to build a model that allows interoperability at a semantic level in EHR. In this way, the model aims to absorb the main difficulties identified in the formalization and structuring of non-structured data from different institutions; the definition and delivery of these in a standard that allows interoperability to minimize barriers to integration between health systems. Figure 10 shows the data processing stages defined in the proposed model and demonstrates what input format each component delivers.

4.2.2.1 Pre-normalization

The first phase has one goal, transforming received data into a common format and adopting this standard format for subsequent processing. Besides, for the training stage of algorithm next layer, we need anonymize the data received (when they still are not). This data comes from the real world, often as raw data, and needs anonymization because it usually comes from hospital repositories. As reinforced by Zuo et al. (2021), raw data is commonly anonymized even for research purposes, with risk assessment for re-identification and utility.

Figure 10: The six stages of transformation data through the information extraction layer.



Source: Created by the author.

In the subsequent processing steps, we adopted at the model the CSV [9] format, the primary format sent by hospitals, followed by the JSON format. Working with non-structured textual entries, a format such as CSV does not interfere with subsequent processing and maintains the original structure. Besides, this model phase does not store the information extraction algorithms in databases. To that end, we defined CSV as the default format to facilitate the process of sending data to the hospitals.

Therefore, in the proposed model, before any use for research purposes, the data must be anonymized and then exported to the target dataset, and preprocessing—to training and test. As our research perspective works with manual entries patient history independently, it is essential to maintain a certain level of traceability regarding the relationship of entries to each other, but utterly patient identification anonymize.

In this way, we intend to anonymize data, which means replacing personal data by codes, and holding the linking between independent entries. Removing all patient identifiers not to reestablish the linking with the persona, only between entries through the codes.

In Table 16 we show five ways to anonymize data. In the model, we defined to apply anonymization to replace personal and sensitive data with code. That method allows to protect patient identification but still hold the traceability from the independent entries using known internal codes. Traceability maintains the possibility to fetch the context of manual entries. In this way, we can check the context of long sentences, even if separated, to promote the relation extraction in cross-sentence.

The adoption of an anonymization method protects patient information and responsible professionals. The application of the technique involves algorithms for locating and replacing personal and sensitive data with synthetic data. In this sense, we evaluated the data received and identified which personal information is part of the manual entries made by the responsible health professionals. Finally, we generated a list

[9] Comma-separated values.

Table 16: Definition of some data protection methods.

Data Protection	**Description**
Anonymization	Irreversible removal of the link between the individual and his or her medical record data at the level that it would be virtually impossible to reestablish the link.
De-identification	Removal or replacement of personal identifiers so that would be difficult to reestablish a link between the individual and his or her data.
Despersonalization	Process of identifying and separating personal from other data.
Exclusion	Prohibition of specific data elements.
Pseudonymization	Identification data is transformed and then replaced by a specifier that cannot be associated with the data without knowing a certain key.

Source: Kushida et al. (2012).

with synthetic data to allow the replacement. The list contains the following data:

- name and last name of patient.
- name and last name of healthcare professional.
- number to the Brazilian identity card (RG).
- number to the Brazilian individual identification (CPF).
- name and last name of professional.
- local (address, health institution).

According to the data types classification covered by the LGPD[10], since the personal identification is removed and modified in the data processing, making it impossible to identify the person and ensure untying, the LGPD does not apply.

4.2.2.2 Preprossesing data

This phase provided the first structure-related support for the non-structured data. In this phase, the model applies the four tasks defined for NLP, which precede the use of data by the learning algorithms.

For the record, the preprocessing step represents intermediate processing before learning algorithms (Named Entity Recognition, Relation Extraction, and Named Entity Disambiguation tasks) on the data. The primary basic preprocessing tasks are

[10]General Data Protection Law (Lei Geral de Proteção de Dados Pessoais, in portuguese) <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>.

minimizing misspelling problems, cleaning up data, removing noise, normalizing expressions, avoiding semantic relevance, and medical means. Only grammatical adjustments.

a) Removal noise and Spelling correction

From raw data it means any metadata, file headers or footers, symbols from file formats or from within databases patterns that come from the health system automatic exported tools. Besides, the available raw data presents misspellings certainly because the manual entries. Spelling correction avoids redundancy once different word representations create different results to the same meaning.

b) Text standardization

Essentially, it deals with abbreviations for recurring terms, presenting itself as one of the significant challenges. The use of terminologies in the health institutions system aims to mitigate abbreviations and acronyms by adopting international vocabularies.

In the data available for research, several terms and abbreviations require manual annotation, as they have specific characteristics in each hospital, and sometimes because each health professional performed the record differently. This phase normalizes the different terms adopted (when straight-identified).

c) Text normalization

In this research, data come from hospital repositories, and usually raw data has particular characteristics. For example, duplicate entries often occur in the data received, both with the same information as partially duplicated information. However, in the end, have new data and update the patient's status. In these cases, the entry has new information; thus, it treats as part of text normalization task. In this sense, normalization at the sentence level, to clear duplicated entries. Also, at the word level, to apply basic NLP tasks, (1) stemming and (2) lemmatization.

In addition, our scenario applies to data in Portuguese, with few tools available and challenges regarding datasets in the health area that allow exploration and training. In this way, we need to promote annotated data for training and adapt tools initially proposed for use in English (the predominant language for NLP tools) to Portuguese. The next component executes that activity to information extraction.

4.2.2.3 Information Extraction

In this phase, the model applies techniques for extracting information, with the objective of (1) localizing and recognizing entities; (2) based on the first outcome,

extracting the relations between the recognized entities; (3) exploring the use of data representations in graphs to allow Named Entity Disambiguation (also called Linking Sentence) categorized with more than one concept.

a) Data annotation

Data annotation is not the context for extracting information but is related, as it is a vital part of the process. According to Lopes, Teixeira and Oliveira (2019), to create a machine learning model that performs tasks such as named entity recognition, we need a collection of annotated texts used both in training and testing the data.

On the one hand, we employed the labeled data made in Portuguese from Portugal, clinical notes, one of the contributions of the work of Portuguese Clinical NER[11]. On the other hand, some characteristics of the labeled data need annotations around the Brazilian context in health records. In this sense, we annotated the clinical notes of hospitalized COVID-19 patients for this research, and more details are explored in Chapter 5.

b) Named Entity Recognition

In this phase, we adopted the proposal of the Transformer architecture, BERT, a Bidirectional Encoder (trained language model with open context data according to Devlin et al. (2018)), and later applied a fine-tuning technique to maximize the reach of the data vocabulary once we brought in a domain. In this way, the objective is to develop a model to explore data in health to improve results in the recognition and extraction of relationships since the vocabularies were updated and specialized for the health domain.

In practical terms, the Named Entity Recognition task refers to everything with a proper name, such as a person, location, and organization. Then, it usually extends the meaning to absorb a range of time, date, and numeric information (identity card and other personal documents). This research extends to an exclusive health record domain to identify these specific entities.

The authors in Vretinaris et al. (2021) discuss this problem using the example with a healthcare staff, where someone can use the "renal disorder" and "kidney disease" in the patient' record to refer to the entity that is defined as "nephrosis" in the patient' records. Besides, another scenario, where the patient records contain an acronym like ARF, and that entity could refer to "acute renal failure" or "acute respiratory failure" in a medical context—some practical examples from real-world data. Using a semantic structure as ontologies, we can enrich classes with another definitions and synons to match medical knowledge from contextualized data joint the entities extracted.

[11]PortugueseClinicalNER <https://github.com/fabioacl/PortugueseClinicalNER>

The Entity recognition and extraction algorithms present output data in a structured format, usually JSON. These data contain the entities recognized from the methods employed. This extraction phase exhausts the techniques to identify and extract information and entities using the learning algorithms. We defined JSON format as the output format for the Structural layer, which was used to exchange data to the Semantic layer.

4.2.3 Semantic layer

At this phase, the processed data already has a semi-structured format, as defined in the last layer, JSON format. The previous phase identified typical terms in the data, information from the dataset to training the learning algorithms provided by the hospitals. However, we seek to standardize the data at international acceptance levels, adopting health standards and vocabularies. For this purpose, recognized entities need representation in international vocabularies, such as SNOMED-CT, ICD-10, ICD-9, and other indispensable classifications.

In this sense, this layer intends to process data to identify entities and link them to standardized concepts. The last layers recognized information nature related to general contexts, such as people, places, and dates, moreover, following applying a NER trained from the healthcare domain, toward a specific recognition, searching for medications, prescriptions, signs, and symptoms, to structure data to adopt a standard effectively. Therefore, the model recognizes the entities it can represent in controlled terms in the semantic layer, defined as the **Entity Normalization** process. Normalization represents the process of, given the recognized entities, identifying which ones relate to controlled vocabularies of the healthcare area and updating them to this new term. In this sense, this layer aim to replace proprietary vocabulary with an internationally accepted vocabulary, using controlled vocabularies as training resources to identify entities, since the proprietary terms generally have particular characteristics to be recognized by more comprehensive techniques such as NER and RE Li et al. (2019).

It is important to note that we followed the recommendations and terminology currently defined as the standard for Brazil through RNDS definition and used in the HL7 FHIR standard, as published on the official RNDS resource site[12]. The vocabularies definitions are independent of the type of standard adopted. Although some standards prioritize using specific vocabularies, this does not happen in the openEHR and HL7 standards, adapted as needed.

Furthermore, in the **Knowledge repository** component stored all terminologies most used in the known healthcare domain. That repository provides a knowledge base to the entity normalization task, besides holding the structured data in the built database when

[12]SIMPLIFIER.NET/RNDS <https://simplifier.net/NationalHealthDataNetwork/>

done the process.

4.2.3.1 Harmonization of standards

In the previous layers, the model treats the data to structuring, relating, contextualizing—techniques for extracting and disambiguation entities besides methods for associating concepts with vocabularies and terminologies. Nevertheless, the level of semantic representation still lacks a standardized format focused on exchanging data between institutions. Adopting a standard for information exchange ensures that data can be sent (owner institution) and received (receiver institution) through the systems, maintaining the meaning of the information allowing full use by the recipient.

Therefore, we still defined a **Harmonization of standards** component at the Semantic layer. It aims to perform a normalization for semantic adjustments to one final healthcare standard. This component prioritizes the definition of new semantic conventions, advocating the sharing and exchange of data adopting a standard. Concerning healthcare standards, there is no global consensus regarding what to adopt. The model uses an intermediary ontology structure to represent data. This approach is well-acceptance in recent studies, once it is possible to define complete information and vocabularies links before selecting a final healthcare standard.

Also, the component in harmonization layer boundary, **Integration method**, provides a communication interface to deliver data to the healthcare institutions, for example, an API interface. Using that interface, the model aims to allow data requests since the data acquisition is performed exclusively by the **Data acquisition method** of the model. Until this point, the model processed data and structured them into the intermediary ontology, and this layer aim to convert to a healthcare standard, such as openEHR, HL7, and others.

4.2.4 Organization layer

The last phase of the proposed model involves policies for data governance, how and who can access the shared data, defining the security levels, and the internal processes to capture and share it. Likewise, the definition of process and workflows to enable the full consent to share data–from the healthcare professionals and the whole institution since the data laws require. In the model, the organization layer represents the interaction point among the healthcare ecosystem. At this stage, structured data allows the sharing and receipt of information with quality.

However, the organizational layer involves fewer technical actions; that phase involves more institutional procedures (from the healthcare institution and system providers) under the management aspect to achieve interoperability. In line with this

vision, we can highlight awareness-raising measures in the institution to adopt terminology effectively, minimize the manual record of textual entries, enable the systems to use health standards for information sharing. Therefore, measures require the institution's commitment, health systems providers, and healthcare professionals involved.

4.3 Final remarks and model contributions

The presented model proposes an integrated view to achieving semantic interoperability in electronic health records. We expect that to promote an integrated scenario, and the model can process hospital data with real-world characteristics in raw format to a structured layout, making it possible to establish new applications, such as interoperability between institutions, ensuring information delivery at a semantic level.

Likewise, once the data is processed and structured, new approaches can be explored, such as using these data for secondary research, mapping new guidelines for public health, and understanding the population's needs at micro-territorial and national levels. Besides, a structured data scenario allows the implementation of an accurate integrated medical record, where health institutions have adequate access to existing patient information from other institutions—and rely on quality data for clinical decision-making as soon as this integrated scenario. The sharing of quality data in clinical research becomes another viable opportunity.

In this scenario, building a model concerned with solving interoperability problems, observing the characteristics within the healthcare institutions, taking their needs as a starting point for the research makes its applicability viable and potential for use after its conclusion, which means practical feasibility.

We highlighted as the model contributions the following: (1) Propose a conceptual model to allow interoperability at the semantic level, exploring different health standard and their feasibility. The currently growing around the quality of health data, the use of controlled terminologies and vocabularies, and maintenance of the security and privacy patient data (BLOOMROSEN; BERNER et al., 2020) are some reasons that stimulated different works to promote interoperability in EHR (LEE; KIM; LEE, 2021; GAGALOVA et al., 2020; ROEHRS et al., 2018).

A semantic interoperability model enables the secondary use of the patients' data once it ensures the capability of the receiving system to understand the meaning of the transmitted data, which plays an essential role in the EHR environment. As reinforced by Adel et al. (2019), semantic interoperability involves structuring data, adopting health standards to allow semantic representations, and including vocabularies and terminologies.

Desiring an interoperable environment, (2) the model defines a set of techniques to structure non-structured data, looking for the real-world scenario. There are many

challenges when a model works with data from the real world using NLP methods. The main aim of different NLP methods is to get a human-like understanding of the text. It helps to examine the vast amount of non-structured and low-quality text and discover appropriate insights ('NASEEM et al., 2021). Whereas using free texts collected by clinicians, nurses, and caregivers can represent a challenging format, it is an opportunity to explore valuable insights that structured – quantitative and tabular – data cannot provide (KORMILITZIN et al., 2021; LAPARRA et al., 2021).

The proposed model aims to provide an interoperable model at the semantic level, focusing on information extraction to explore non-structured data. Moreover, exploring the potential approach of NLP to maximize the value of the EHR and healthcare data makes data a critical and trusted component in improving health outcomes (SIVARETHINAMOHAN; SUJATHA; BISWAS, 2021). According to Kormilitzin et al. (2021), identifying medical concepts and information extraction is a challenging task, yet an essential ingredient for parsing unstructured data into structured format. The model normalizes raw data acquired to use the Named Entity Recognition task and extracts the recognized entities (PERERA; DEHMER; EMMERT-STREIB, 2020).

The ambiguity and variability problems in clinical concepts, mainly in scenarios originating from unstructured data, usually emerge in information extraction approaches, and we treated them in a post-data processing step, over the ontology structure. According to Newman-Griffis et al. (2021), ambiguity consists of words and sentences with different concepts depending on their context, a problem extensively researched jointly with extracting information from the biomedical domain. In this sense, from the recognized entities and later the relationships extracted from the unstructured data into our ontology structure, we assigned each entity to a correct mention in a knowledge base that can be linked to validated ontologies of COVID-19, which helps to consolidate the semantic context.

In this way, the proposed model experimented with information extraction techniques, and (4) proposed an improvement in the best-evaluated model to named entity recognition. The proposed model has a set of information extraction techniques, including experiment different NER models to compare the most effective outcomes and fine-tune the best model using the annotated data. The authors in Kormilitzin et al. (2021), Schneider et al. (2020) had demonstrated an improvement to NER tasks applying a fine-tune on a small sample from the target domain dataset. They reinforced that despite a close similarity between the data sets and the NER tasks, the outcomes improved from a fine-tuning of the target domain.

Furthermore, information extraction models need high-quality annotated data in the healthcare domain. Regarding that, and focusing on the annotated corpus Portuguese lacks, we proposed contributing with an annotated dataset. Thereby, to contribute to the model performance, (5) this research aims to develop an annotation methodology

data. As discussed by the authors Kormilitzin et al. (2021), although we have witnessed an advance in the performance of models for extracting information, NLP applied to the health domain. There is still a significant challenge regarding reliable models for extracting clinical data due to the scarcity of high-quality annotated examples (observing the area in general) and the lack of Portuguese annotated tools and corpora.

The authors in Lopes, Teixeira and Oliveira (2019), Oliveira et al. (2020) had brought significant contributions in this regard, with data annotated in Portuguese, and the work of Schneider et al. (2020) shows BioBertPT[13] model with excellent outcomes to NER task using those Portuguese annotated corpora. In this sense, we propose for this research to annotate a dataset COVID-19 in Portuguese. We propose to conduct a workshop first to demonstrate the use of tools and vocabularies frequently used and explain the methodology of the chosen software. Likewise, to empower a team to collaborate in the annotation and verification process of data. As reinforced by Nasar, Jaffry and Malik (2021), novel research has been exploring NER in non-English languages, although the annotated training data still may have a high cost to generate (HOUSSEIN; MOHAMED; ALI, 2021).

We emphasize below the technical characteristics of the contributions of this model, which together form a comprehensive solution for improving the processing and interoperability of unstructured health records in Portuguese:

An integrated conceptual model to extract, structure, and enrich non-structured health records: This contribution addresses the pressing need to deal with the large volume of unstructured data in healthcare, where critical information is often trapped in free text. By integrating unstructured clinical notes, the model extracts relevant clinical entities and enhances the data for subsequent use, ensuring higher quality and better interoperability across systems. Our work introduces a novel approach by evaluating three Transformers models, including BERT, on the healthcare domain and its variants, specifically for the Portuguese language. This evaluation is a significant step forward, as most NLP models are trained in English. Adapting these models to Portuguese demonstrates their applicability and effectiveness in languages with fewer resources, which was interesting in our research.

A fine-tuned model trained on a small specialized COVID-19 dataset in Portuguese clinical notes: Despite the challenges of working with a small dataset, the fine-tuned model achieved high precision in extracting entities related to COVID-19. It shows the model's ability to generalize despite limited data, making it a valuable tool for extracting structured information from unstructured clinical records in Portuguese.

A lexical dictionary to normalize entities to guide entity linking initiatives: A crucial challenge in NLP is normalizing entities, where different terms refer to the same concept.

[13]BioBERTpt - Portuguese Clinical and Biomedical BERT <https://github.com/HAILab-PUCPR/BioBERTpt>

Developing a lexical dictionary allows for the standardization of extracted entities, ensuring that different terms used by healthcare professionals can be linked to a common reference, enhancing interoperability and facilitating more accurate data analysis.

We provide a robust methodology for annotating clinical data, a key technical contribution, particularly in Portuguese, where annotated datasets are scarce. This contribution not only supports the development of machine learning models but also instills confidence in the reliability of our research, serving as a valuable resource for future research in the field of healthcare NLP.

A study and explanation about a regional COVID-19 clinical notes dataset: Analyzing a regional dataset from Porto Alegre provides valuable insights into the unique characteristics of clinical practices in South Brazilian hospitals during the COVID-19 pandemic. This context-specific study adds depth to understanding how data is recorded and structured in real-world clinical environments.

Our research makes a significant impact on healthcare standards and the health domain by proposing a semantic interoperability model for unstructured clinical notes in Portuguese. This model directly addresses the gap in structured data in Portuguese healthcare systems, aligning extracted data with international healthcare standards and ontologies (e.g., COVID-19-specific ontologies). The model's value lies in its ability to ensure that data from unstructured clinical notes can be integrated into broader healthcare systems, facilitating more effective information exchange and continuity of care.

5 MATERIALS AND METHODS

This chapter presents the adopted methodology, focusing on implementing the necessary components to evaluate the proposed model in Chapter 4. We have considered recent techniques in information extraction, emphasizing the tasks of named entity recognition and lexical normalization. The implementation applied state-of-the-art technologies, such as Bidirectional Encoder Representations from Transformers (BERT) architectures, to evaluate the conceptual model. We used OWL ontology to represent the extracted data.

The experiments and data processing pipeline, including the proposed interoperability model through a prototype, are summarized in Figure 11. The experiments aimed to evaluate the conceptual model and assess its feasibility. We outlined three steps to map the behavior and viability of the components, followed by a description of the process and technologies envisioned for implementation. The first step consisted of organizing and cleaning the dataset, followed by annotating a small subset to enable the training of data within a specific COVID-19 domain. The second step focused on recognizing and extracting entities, specifically through a Named Entity Recognition (NER) task applied to unstructured data. In the third step, we established the classes composing the ontology structure, ensuring alignment with validated ontologies that represent COVID-19-related data. Finally, we performed a data harmonization process according to healthcare standards, analyzing the context represented in the designed ontology and mapping it to the HL7 FHIR standard.

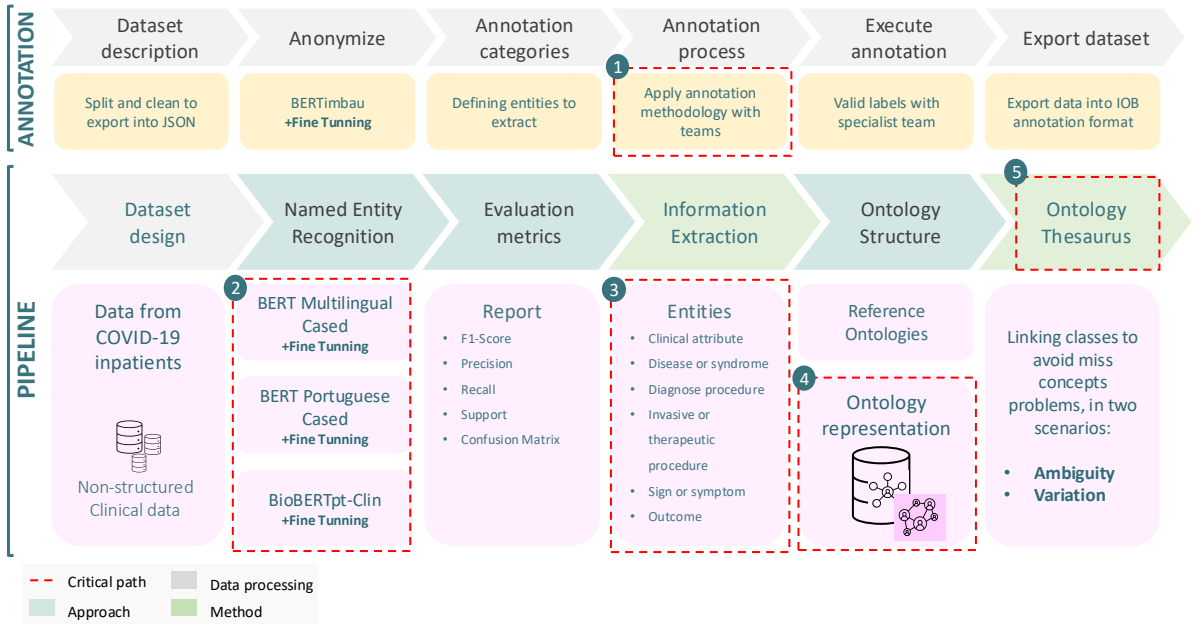
5.1 Annotation

The data used in this research came from hospital partners, focusing on the clinical notes of hospitalized patients who tested positive for COVID-19. In general terms, we needed to define the preliminary information we aim to extract and structure from the raw dataset provided based on the domain scope of COVID-19.

This research has been approved by the Research Ethics Committee (CEP) by number CAAE: 33540520.6.3004.5327 as part of the project titled "Modelo Inteligente de Blockchain para Informações de Saúde e Interação com Pacientes no âmbito da COVID-19," referred to MyDigitalHealth project. The project was submitted to Platform Brazil and approved by the Grupo Hospitalar Conceição (GHC)[1], Santa Casa de Misericórdia (ISMCPA)[2], Hospital Ernesto Dornelles (HED)[3], Hospital de Clínicas de

[1]Grupo Hospitalar Conceição <https://www.ghc.com.br/>
[2]Santa Casa de Misericórdia <https://santacasa.org.br/>
[3]Hospital Ernesto Dornelles <https://www.hed.com.br/>

Figure 11: Designed pipeline to experiment with the proposed interoperability model through a prototype.



Source: Created by the author.

Porto Alegre (HCPA)[4], Hospital Moinhos de Vento (HMV)[5], Hospital Universitário de Santa Maria (HUSM)[6] hospital partners and Universidade do Vale do Rio dos Sinos (UNISINOS).

The research aims to apply the proposed interoperability model to unstructured COVID-19 data. However, to advance in model training and entity extraction, we needed annotated data in the specific context of COVID-19 in Portuguese, a characteristic we did not find in existing models. With this need in mind, we invited medical students from the 6th semester onwards to carry out the annotation process. Additionally, two healthcare specialists (a Ph.D. in Medical Sciences and a Pharmaceutical Biochemist) assisted the students whenever there were doubts regarding clinical concepts. These specialists had access to sensitive data as part of the research project as postdoctoral researchers. Therefore, all data annotated by the student team and reviewed by these specialists were anonymized.

5.1.1 Dataset description

Considering the potential to interoperate data, we reviewed the datasets we received to determine how to apply annotations and select the most suitable one. In a preliminary

[4]Hospital de Clínicas <https://www.hcpa.edu.br/>
[5]Hospital Moinhos de Vento <https://www.hospitalmoinhos.org.br/>
[6]Hospital Universitário de Santa Maria <https://www.gov.br/ebserh/pt-br/hospitais-universitarios/regiao-sul/husm-ufsm>

Table 17: ISMCPA description of the raw dataset

Dataset description	**Count**
Total of columns	84 variables
Total of null-columns	42 variables
Columns more than 80% filled-out	4 variables
Total of rows	631,147 entries
Total rows for clinical notes variable	631,136 entries

Source: Created by the author.

analysis, we found that the **HUSM**, **HCPA**, and **HMV** datasets did not contain clinical notes, instead offering primarily tabular data on signs and symptoms, exams, lab results, and some observations. The **HED** dataset presented inpatient monitoring data from the Intensive Care Unit (ICU), which was also unsuitable for our research. Given our focus on information extraction from clinical notes, these four datasets were excluded from further consideration.

The **ISMCPA dataset** comprised 84 variables, with some columns containing over 631,000 entries. Key columns for our analysis include `clinical_note_code`, which records the evolution code, and `clinical_note_date`, representing the evolution date. Additionally, `clinical_note_type` captures the type of evolution, while `patient_code` serves as the patient identifier. Metrics such as `weight`, `height`, and `body_surface` provide patient physical attributes, though their incomplete data limited their use in annotating clinical notes. The `clinical_note` column contains full-text descriptions of the clinical evolution notes in rich text format (RTF), the variable of interest to the research. In Table 17, we summarized the key variables of the studied dataset.

Several variables are related to managing the EHR system, such as `user_number`, the user's name, and `record_status`, indicating the situation of the record, and many fields contain no data. Despite these inconsistencies, the dataset provides a comprehensive structure for analyzing clinical evolution records. However, we couldn't use the dataset for the research, as it comprises a mix of clinical notes from nurses, specialist doctors, and on-call doctors; each entry follows different structures and characteristics. Nurse notes typically focus on daily care and patient observations, emphasizing the patient's general condition, comfort, and medication administration. In contrast, these notes usually lack detailed diagnostic evaluations, complex exam interpretations, or invasive or therapeutic procedures. Their primary goal is to document the care provided, the patient's response to treatment, and overall physical condition. Since our research focuses specifically on medical clinical notes, we decided to exclude this dataset from the initial phase of the study.

The last analyzed dataset, from **GHC hospital** contains 65,535 records and 13 variables, among which we have `patient_id`, `gender`, `birthdate`, `provider`, `case_id`, `date_encounter`, `primary_diagnostic`, `discharge_status`, `clinical_note_data`, and

Table 18: GHC description of the raw dataset

Dataset description	**Count**
Total of columns	13 variables
Total of null-columns	2 variables
Columns more than 80% filled-out	8 variables
Total of rows	65,535 entries
Total rows for clinical notes variable	65,535 entries

Source: Created by the author.

`clinical_note` are fully populated, each containing non-null values for all 65,535 records. These columns form the core structure of the dataset, providing essential patient demographics (e.g., gender and birthdate), encounter details (e.g., provider and date of encounter), and diagnosis data. Notably, `primary_diagnostic` contains the primary diagnosis information, while `discharge_status` likely categorizes each patient's care outcome. In Table 18, we summarized the variables and key descriptions of the dataset.

On the other hand, some columns exhibit a significant amount of missing data that didn't represent more than 50% of the total data, then they weren't considered. Moreover, the drugs and `lab_tests` columns are completely null, with no recorded data. This suggests that these fields were either not collected. Despite these gaps, the dataset's key fields remain comprehensive, offering valuable use for this research focused on patient encounters, diagnoses, and care outcomes based on medical clinical notes.

Considering our research goals, we evaluated both the **GHC** and **ISMCPA** datasets. Using the ISMCPA dataset would have required filtering to separate medical clinical notes from nurses' notes due to their different data annotation structures. Therefore, we chose the **GHC dataset** for the annotation process and data anonymization. We used this dataset for the remaining annotation experiments, training, and future analyses.

5.1.2 anonymization

We performed a data anonymization process before beginning the studies on the available data, using a BERT model pre-trained on Portuguese language data. We applied this model to classify and identify data such as proper names, organizations, and locations. We chose the BERTimbau model because it is a well-established model for Portuguese, selecting a fine-tuned version on Hugging Face, *ner-bert-base-cased-pt-lenerbr*[7], with nearly 20,000 downloads in January 2022 and 93,343 (ninety-three thousand, three hundred and forty-three) downloads in August 2024. The model was fine-tuned using the LeNER juridic dataset for NER tasks to

[7]<https://huggingface.co/pierreguillou/ner-bert-base-cased-pt-lenerbr>

recognize the generic entities PERSON, LOCATION, and ORGANIZATION. The objective of our research is not to annotate data for generic entities like these; therefore, we used an existing and established model. The proposed model in this research aims to promote the semantic interoperability of unstructured data within the COVID-19 domain, thus focusing on specialized annotations.

In this context, we first processed the data and applied initial anonymization to all potentially sensitive and identifiable information on 1,000 documents. To avoid losing the real structure of the texts, the identified information was replaced with synthetic information, such as new names, organizations, and locations. Personal information, including names, addresses, and other sensitive details recognized in the records by the NER model, we handled to prevent the identification of the patient, the care team involved, and the healthcare institution.

In Figure 12, we brought only the recognized entities/words, and we can understand 6 important pieces of information for processing the result and applying it to clinical documents. The column **entity** represents the identified entity type. Then, column **score** represents the model's confidence in predicting the entity, ranging from 0 to 1. A value close to 1 indicates high confidence in the prediction, and finally **index** represents the position of the token in the input sequence (token index). As a characteristic of the BERT Tokenizer, tokens are fragmented into subtokens called WordPiece Tokenization. This approach was adopted to reduce the number of unknown words by breaking them into subwords or roots. The model can better handle rare or new words by utilizing smaller, known tokens. We can observe this in the **word** column, where the token identified as part of the entity is a word fragment (tokens starting with ## indicate sub-words that were split by the tokenizer). Finally, **start** and **end** represent the starting and ending positions of the token in the original text sample.

Then, looking inside the table values, we can see different tags, like **B-PERSON**, **I-PERSON**, **I-ORGANIZATION**, or **B-ORGANIZATION**. This is a type of annotation format called IOB (Beginning, Inside, Outside), where the prefixes *"B-"* and *"I-"* indicate that an entity is the beginning of the word or expression and the continuation of the entity respectively, (e.g., B-PERSON indicates the beginning of a person's name), and **I-** constitutes the continuation of the entity that began on **B-** (e.g., I-PERSON indicates the continuation of a person's name).

Analyzing the highlighted colored squares around the data on Figure 12, we can see the PERSON entities were recognized in the clinical note from lines 0 (zero) to 7 (seven) (e.g., B-PERSON and I-PERSON), with the first referring to patient identification and the last one, in line 32 (thirty-two), referring to the doctor. Additionally, the ORGANIZATION entity in line 33 (thirty-three) recognized the hospital, and from lines 36 (thirty-six) to 38 (thirty-eight), it identified the name of the hospital. Also, we presented another way to analyze that information, using the same clinical note fragment with the recognized

Figure 12: Recognizing the entities from the patient information.

	entity	score	index	word	start	end
0	B-PESSOA	0.999713	26	B	53	54
1	B-PESSOA	0.999669	27	##LA	54	56
2	B-PESSOA	0.999673	28	##N	56	57
3	B-PESSOA	0.999559	29	##DA	57	59
4	I-PESSOA	0.999863	30	DE	60	62
5	I-PESSOA	0.999869	31	M	63	64
6	I-PESSOA	0.999877	32	##EL	64	66
7	I-PESSOA	0.999883	33	##LO	66	68
8	B-TEMPO	0.985013	40	12	89	91
9	B-TEMPO	0.962862	41	/	91	92
10	B-TEMPO	0.967999	42	03	92	94
11	B-TEMPO	0.812397	43	/	95	96
12	B-TEMPO	0.996948	49	13	112	114
13	B-TEMPO	0.998119	50	/	114	115
14	B-TEMPO	0.997591	51	03	115	117
15	B-TEMPO	0.997553	52	/	118	119
16	B-TEMPO	0.722421	57	03	127	129
17	B-TEMPO	0.998036	109	04	255	257
18	B-TEMPO	0.998399	110	/	257	258
19	B-TEMPO	0.997034	111	03	258	260
20	B-TEMPO	0.998760	118	04	278	280

	entity	score	index	word	start	end
21	B-TEMPO	0.998860	119	/	280	281
22	B-TEMPO	0.998513	120	03	281	283
23	B-TEMPO	0.768403	130	/	305	306
24	B-TEMPO	0.611976	131	03	306	308
25	B-TEMPO	0.857274	137	/	318	319
26	B-TEMPO	0.674583	138	03	319	321
27	B-TEMPO	0.942911	156	/	362	363
28	B-TEMPO	0.657846	157	03	363	365
29	B-TEMPO	0.549033	161	17	374	376
30	B-TEMPO	0.925469	162	/	376	377
31	B-TEMPO	0.884689	163	03	377	379
32	B-PESSOA	0.999484	347	Helena	935	941
33	B-ORGANIZACAO	0.533857	353	Hospital	966	974
34	I-LOCAL	0.748449	354	Ernesto	975	982
35	I-LOCAL	0.603787	355	Do	983	985
36	I-ORGANIZACAO	0.634670	356	##er	985	987
37	I-ORGANIZACAO	0.656664	357	##nel	987	990
38	I-ORGANIZACAO	0.506860	358	##es	990	992

Source: Created by the author.

entities in Figure 13, where we display it in a more user-friendly format, with the subtokens (WordPieces) reconstructed to form the complete word.

Figure 13: Recognizing the entities from the patient information.

PACIENTE:24600 EVOLUCAO: UTI Area 3 - plantao manha B##LA##N##DA DE M##EL##LO, 30 anos Emergencia 12/03 / Semi-intensiva 13/03 / UTI 17/03 LISTA DE PROBLEMAS ALERGIA A À MORFINA HAS Obesidade Atual Pneumonia viral por COVID-19 -PCR SARS-CoV-2 + (04/03) -Sintomas desde 04/03 - Dexametasona DI 13/03 - VNI 13/03 UTI por piora ventilatorio - IOT+VM 17/03 - Prona 17/03 Subjetivo - Ao exame Sedada e bloqueada Estavel hemodinamicamente Pronada ontem a tarde VM peep 12 FR 30 VAC 330 mL Fio2 50% sat 98% - plato 27 DP 15 - p/f 224 ph 7,34 pco2 79 po2 112 hco3 341 Abdomen depressivel DIurese 1500 mL/6h Creat 0,5 ureia 60 na 153 k 4,3 PCR 91->303->341 Melhora relação p/F apos prona - hipercapnia melhor BCP sobreposta em tratamento FUnção renal normal - boa resposta a diuretico - hipercalemia resolvida Quadro global grave Retorno posição supina - prona ha mais de 16 horas Aguarda culturais. Sob os cuidados da Dra. Helena. Atendimento emergencia Hospital Ernesto Do##er##nel##es.

Color legend

- B-PESSOA
- I-PESSOA
- B-TEMPO
- I-LOCAL
- B-ORGANIZACAO
- I-ORGANIZACAO

Source: Created by the author.

Since our goal was to maintain the structure of the clinical notes while de-identifying the patient, we created a spreadsheet of synthetic names to use. As shown in Figure 14, we applied the replace process only on the recognized entities, replacing those related to

Figure 14: Fragment of the developed script to replace the recognized patient information with synthetic data.

```python
def deidentify_note(note, entities, shuffled_names, shuffled_orgs, shuffled_places):
    """
        Arg:
            entities (list): consolidated entities as B- and I- into the main entity
                             (e.g.: B-PESSOA and I-PESSOA = PESSOA)
    """
    # replace entities to synthetic data
    for entity in entities:
        if entity['entity_group'] == "PESSOA":
            note = replace_text(note, entity['word'], random_choice(shuffled_names))
        elif entity['entity_group'] == "ORGANIZACAO":
            note = replace_text(note, entity['word'], random_choice(shuffled_orgs))
        elif entity['entity_group'] == "LOCAL":
            note = replace_text(note, entity['word'], random_choice(shuffled_places))

    return note
```

Source: Created by the author.

PERSON[8], LOCALIZATION[9], and ORGANIZATION[10] with synthetic information.

After anonymizing the information, considering the data were unstructured and contained sensitive information, we completed the anonymization step and exported each row from the CSV file into individual Plain Text files. It shows the size variability of each clinical note reflecting a real-world characteristic; there are very small and bigger notes. Then, we conducted a qualitative validation with the specialists involved in the project, who were approved by the ethics committee to access the data, to ensure that the documents were fully anonymized. We are aware that manual evaluation requires significant effort from the team. Still, since the textual data presented a variable structure, we encountered names with abbreviations that could indicate a correlation with a specific patient if a thorough study were conducted. For this reason, we manually validated the documents and replaced any information that could last to patient identification.

5.1.3 Annotation categories

After completing the data anonymization step, we analyzed the dataset to identify possible entities of interest related to the COVID-19 domain, which is essential for patient continuity of care. The categories were discussed in collaboration with medical students and specialists involved in the further data annotation process. While the first entity definitions were based on previous studies, they were adapted to fit the specific scenarios of our clinical notes of COVID-19, in Table 19, we presented the Named Entities defined

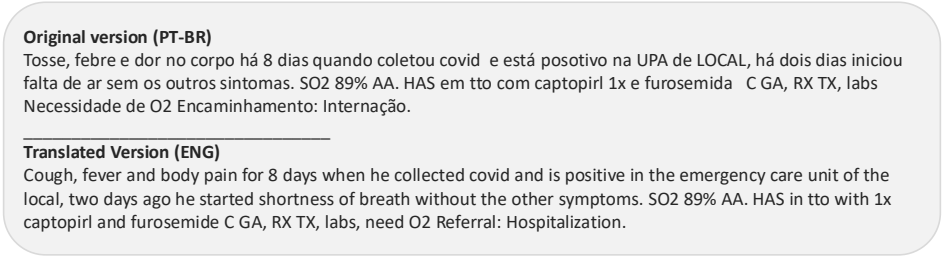
[8]PERSON in Portuguese PESSOA
[9]LOCALIZATION in Portuguese LOCAL
[10]ORGANIZATION in Portuguese ORGANIZAÇÃO

Table 19: Entities defined to data annotation phase based on the clinical notes received.

Entity (PT-BR)	**Entity (ENG))**
Atributo Clinicos	Clinical Attributes
Resultado	Result
Doenca ou Sindrome	Disease or Syndrome
Sinal ou Sintoma	Sign or Symptom
Procedimento Diagnostico	Diagnosis procedure
Procedimentos Invasivo ou Terapeutico	Invasive or Therapeutic Procedure

Source: Created by the author.

Figure 15: Example 1 of a clinical note received to build the dataset on the COVID-19 domain.



Source: Created by the author based on the data received.

to create the dataset.

In Figures 15 and 16, we showed an example of the clinical notes format received.

We converted it into JSON format and loaded it into a dedicated tool to begin the annotation activities, making the annotation process interactive for the team. The annotation classes were partially defined in a generic context, such as disease, chemicals and drugs, medical procedure, diagnostic procedure, disorder or syndrome, test or laboratory result, medical device, pharmacological substance, quantitative concept, signs or symptoms, therapies, among others, as needed to support the text of the clinical notes as a whole. The annotation process is interactive, where each document is opened by the annotators, with the view as shown in Figure 17. Then, the annotations can systematically identify and annotate the entities on the document.

We highlight that the anonymization was performed before entering the data into the UBIAI annotation system to prevent direct identification of the patient or correlation with other clinical notes that may exist for the same patient. It is important to note that the focus of the evolutions is not on analyzing temporal data from free texts but rather on identifying terms and their structuring. Therefore, we used the traditional method of irreversible anonymization on our data. Since the presence or absence of data influences the study's objective, and the organization pattern of ideas, sentence type, and structure must be recognized for information identification and extraction, a dictionary of

Figure 16: Example 2 of a clinical note received to build the dataset on the COVID-19 domain.

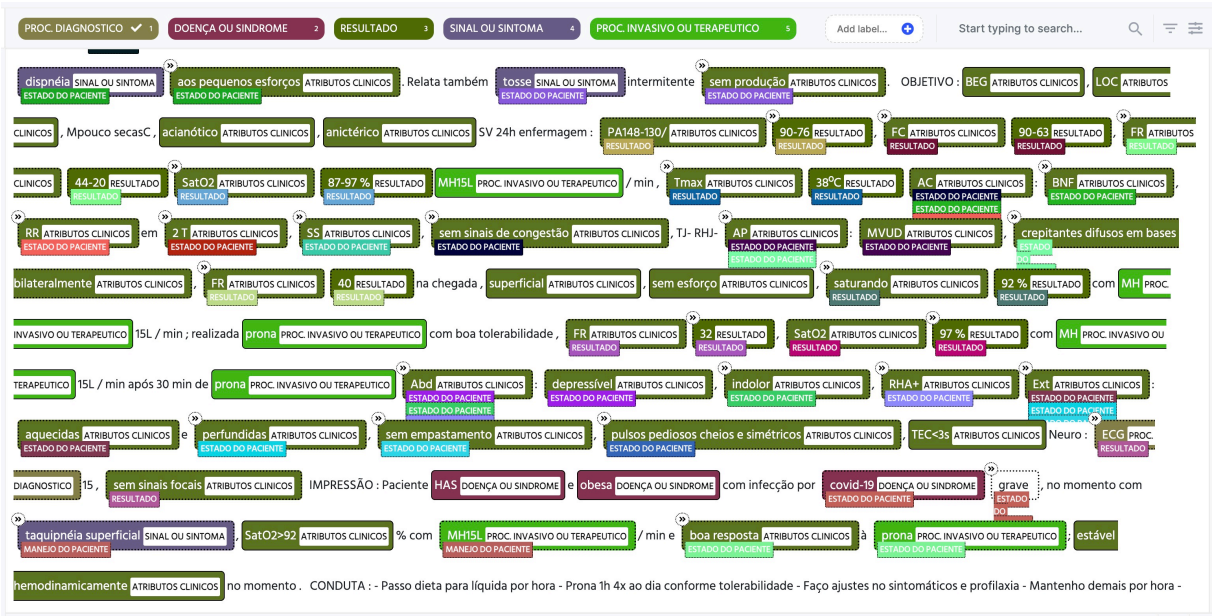
Original version (PT-BR)
PLANTÃO COVID Angela Silva Brasil, 31 anos Chamado por dessaturação e taquipnéia. Paciente tranquila, refere dispnéia após ir ao banheiro. Sintomática há mais de sete dias. Ao exame, REG SatO2 92\% CN 6l/min - FR 24 Sem esforço ventilatório Taquipnéia superficial AR MV+ difusamente diminuido sem RA I- SRAG por covid-16 sem sinais de esforço ventilatório C- MH 8l/min, titular litragem de O2 conforme SatO2 Salbutamol inalatório Oriento decúbito lateral ou posição prona Sem mais acréscimos por ora.

_____

Translated Version (ENG)
COVID SHIFT Angela Silva Brasil, 31 years old. Called due to desaturation and tachypnea. Calm patient, reports dyspnea after going to the bathroom. Symptomatic for more than seven days. On examination: REG SatO2 92% on CN at 6 L/min, FR 24, no respiratory effort. Superficial tachypnea. Lungs: AR MV+ diffusely diminished, no RA. I- SARS due to COVID-19, without signs of respiratory effort. C- MH at 8 L/min, titrate O2 according to SatO2. Inhaled salbutamol. Recommend lateral decubitus or prone position. No further additions for now.

Source: Created by the author based on the data received.

Figure 17: Interface of the annotation tool used by the annotation team to manually handle the task on the clinical notes.



Source: Created by the author .

personal data containing synthetic information was used to replace the existing personal information. The process involved random selection of information, with no possibility of establishing relationships across the evolutions, even if the same service, patient, or physician receives different synthetic information.

5.1.4 Annotation process

Our annotation process was a collaborative effort that included medical students from the 6th semester, with a significant step towards defining the contexts relevant to the continuity of care within COVID-19. This collaboration, which involved students from multiple universities, was a collective effort. We needed to define a methodology that could be replicated and later consolidate the data to formally the final specialized Dataset on COVID-19.

We structured a document detailing the methodology, presenting the project, its objectives to annotate a small and specialized dataset in the COVID-19 context, the fundamental articles supporting the research, and expectations for volume. Together with the annotators, we analyzed the data to define the tags that would be most appropriate and important for capturing the terms contained in the clinical notes. We conducted an annotation test, applying all the defined classes to the selected data and verifying whether the relevant terms were covered. Once the annotations had matured and we realized that we could extract the most relevant information from the clinical notes, we brought the classes to healthcare professionals for discussion and validation, receiving affirmative feedback.

The annotation team continuously performed these iterative steps, classifying medical terms and expressions in the clinical notes into generic classes relevant to the COVID-19 domain. By using well-adjusted generic classes across different free-text structures, we ensured better accuracy in training the machine learning models, as this allows the generalization of language patterns across different types of data. However, using only generic classes in a specific domain context, such as this research on COVID-19, may fail to recognize essential terms due to the absence of these data in the training process. Once the annotations were completed, we trained and experimented with the selected machine-learning models.

With the annotation process complete, we could define the knowledge scope for the ontological structure we created at the end of the project to represent the extracted information. This ontology is presented in more detail in section 6.5. We understand the limitations and challenges of defining an ontology in healthcare, especially COVID-19. A COVID-19 case can progress to a wide range of symptoms, from asymptomatic to more severe conditions. Therefore, based on the domain-specific annotations, it was possible to develop strategies to identify and mitigate scenarios where COVID-19 cases evolve while

limiting the representation to the context of COVID-19. Finally, the data extracted by the machine learning models can be represented within the defined ontology, ensuring an intermediary data structure compatible with conversion to any preferred healthcare standard, such as openEHR and HL7 FHIR, among others.

5.1.5 Annotation execution

Although the annotation of databases by humans and experts always involves a time-consuming and costly process in projects, in cases where the application is directed toward sensitive areas, such as healthcare, working with semi-automatic or automatic annotations becomes concerning and, at times, complex. In such cases, in addition to the annotation process, the validation is conducted by healthcare professionals.

For this study, we applied a distributed team annotation methodology to maximize the number of annotated and qualified documents by employing 6th-semester medical students to collaborate in annotating small data corpora. This approach yielded promising results, particularly when the goal is the exploration of specific datasets, as in the current research focused on databases of hospitalized patients who tested positive for COVID-19.

In this scenario, developing and proposing a methodology for annotating small datasets is quite valuable, given the good results that a small annotated corpus can bring when applying specialized language models, such as fine-tuning downstream tasks on BERT for the COVID-19 context.

Our study has demonstrated the effectiveness of a collaborative and distributed annotation methodology. This approach, which involved medical students in the intermediate stages of their course, proved to be a valuable tool in qualifying small corpora. Our research focused on clinical data from hospitalized patients with COVID-19. Despite the significant challenges posed by manual annotation processes, such as high time and resource costs, the collaboration between experts and students played a crucial role in enabling greater annotation efficiency and quality, particularly in the sensitive healthcare domain. This methodology has the potential to significantly improve the quality and efficiency of small dataset annotation, particularly in healthcare research. We organized the steps as follows to execute this methodology.

- Division of annotation teams between Feevale and Unisinos.
- Training of annotation teams.
- Selection of appropriate annotation software.
- Detailed study of the database.
- Definition of the entities to be annotated.

Table 20: Examples of domain, local, or regional acronyms and their respective description.

Clinical word	**Original (PT-BR)**	**Translated (ENG)**
IRH	Insulina Regular Humana	Human Regular Insulin
IOT	Intubação oreotraqueal	Orotracheal intubation
VM	Ventilação Mecânica	Mechanical Ventilation

Source: Created by the author.

- Reference to articles for understanding the defined entities.
- Creation of a guiding document for the use of entities.
- Regular meetings to clarify doubts.
- General review of annotated documents.

5.1.6 Dataset exportation

The annotation process was finalized, and we prepared the dataset for further training. The UBIAI annotation tool, allow us to selecting between many formats to export the data. We chose to export them in the IOB-POS (Beginning, Inside, Outside-Part Of Speech) format, which allows us to work with the annotations in three columns, as shown in Figure 18.

In the Figure 18, the first column represents the word from the clinical note, the second column represents the POS information, and the third column represents our IOB annotation. We presented a common issue on clinical unstructured datasets in Table 20. Health acronyms in the clinical notes rely on regional, local, and proprietary expressions (e.g., within the hospital), and over a small and specialized dataset, it substantially impacts the entity recognition process.

Once the dataset was finalized, we organized the information extraction using the annotation tool UBIAI, which allows downloads in various formats. With this step completed, we began the experiments proposed in the evaluation pipeline, ensuring that the data was ready for processing, entity extraction, and other tests planned to validate our semantic interoperability model. Thus, we completed converting essential information from the raw unstructured data into structured data in the proposed ontology.

5.2 Pipeline considerations

The data annotation process was necessary to support our Named Entity Recognition (NER) experiments. We only performed the information extraction experiments after

Figure 18: Fragment of an IOB file representing some of the entities we defined to the specific COVID-19 domain.

SRAG	NN	B-DOENÇA OU SINDROME
-	:	O
COVID	NN	B-DOENÇA OU SINDROME
19	CD	I-DOENÇA OU SINDROME
ATB	NN	O
:	:	O
Amox	NN	B-PROC. INVASIVO OU TERAPEUTICO
/	NN	I-PROC. INVASIVO OU TERAPEUTICO
Clav	NN	I-PROC. INVASIVO OU TERAPEUTICO
17-19/04	JJ	O
Azitro	NNP	B-PROC. INVASIVO OU TERAPEUTICO
17/04	CD	O
Pipetazo	NN	B-PROC. INVASIVO OU TERAPEUTICO
19/04	CD	O
atual	JJ	O
Vancomicina	NNP	B-PROC. INVASIVO OU TERAPEUTICO
20/04	CD	O
atual	JJ	O
S	NN	O
-	:	O
O	NN	O
-	:	O
MEG	NN	B-ATRIBUTOS CLINICOS
,	,	O
sedado	NN	B-PROC. INVASIVO OU TERAPEUTICO
e	NN	O
curarizado	NN	B-PROC. INVASIVO OU TERAPEUTICO
,	,	O
corado	NN	B-ATRIBUTOS CLINICOS
,	,	O
anictérico	NN	B-ATRIBUTOS CLINICOS
,	,	O
com	NN	O
TOT	NN	B-PROC. INVASIVO OU TERAPEUTICO
em	NN	O
VM	NN	B-PROC. INVASIVO OU TERAPEUTICO
flo2=55	NN	B-PROC. INVASIVO OU TERAPEUTICO
%	NN	I-PROC. INVASIVO OU TERAPEUTICO
e	NN	O
afebril	NN	B-ATRIBUTOS CLINICOS
.	.	O
PA	NN	B-ATRIBUTOS CLINICOS
=	NN	O
164	CD	B-RESULTADO
/	NN	I-RESULTADO
70	CD	I-RESULTADO
FC	NN	B-ATRIBUTOS CLINICOS

Source: Created by the author.

Figure 19: Reprocessed annotation converting into sentences and respective word labels.

	sentence	word_labels
0	PACIENTE:329320 EVOLUCAO : UTI Área 3 - Rotina...	O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,...
1	PACIENTE:24600 EVOLUCAO : UTI ÁREA 3 - ROTINA ...	O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,...
2	PACIENTE:104019 EVOLUCAO : ISOLAMENTO COVID-19...	O,O,O,O,B-DOENÇA OU SINDROME,B-ATRIBUTOS CLINI...
3	PACIENTE:329320 EVOLUCAO : UTI Área 3 - Plantã...	O,O,O,O,O,O,O,O,O,O,O,O,B-ATRIBUTOS CLINICOS...
4	PACIENTE:329320 EVOLUCAO : UTI Área 3 - plantã...	O,O,O,O,O,O,O,O,O,O,O,O,B-ATRIBUTOS CLINICOS...
...	...	...
309	PACIENTE:27650 EVOLUCAO : Sala Veremelha - COV...	O,O,O,O,O,O,B-DOENÇA OU SINDROME,O,O,O,O,O,O,O...
310	PACIENTE:27650 EVOLUCAO : Sala Vermelha COVID ...	O,O,O,O,O,B-DOENÇA OU SINDROME,O,O,O,O,O,O,B...
311	PACIENTE:27650 EVOLUCAO : Sala Veremelha - COV...	O,O,O,O,O,O,B-DOENÇA OU SINDROME,O,O,O,O,O,O,O...
312	PACIENTE:24600 EVOLUCAO : UTI ÁREA 3 - ROTINA ...	O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,...
313	PACIENTE:24600 EVOLUCAO : UTI Plantao area 3 [...	O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,O,...

314 rows × 2 columns

Source: Created by the author.

completing the data annotation phase. An essential step in an annotation methodology is monitoring the quality of the annotations, especially in a small and specialized dataset, as well as the potential for data balancing.

5.2.1 Dataset design

We began the experiments by processing the files exported from the UBIAI annotation tool, as many documents generated with the CSV extension had formatting issues. For this reason, we switched to exporting them in TSV (Tab-Separated Values) format. We then performed preliminary processing to repair the files. After that, we loaded the annotated data into our Jupyter Notebook and pre-processed the shape of the dataset the way we needed to input data to execute the extraction experiments. In the first step, we developed the script to read the TSV file, removed the POS column, and consolidated the annotated words into complete sentences, concatenating by lines and the respective IOB tags in a separate column in the same way. In Figure 19, we show the final format of annotations with their corresponding tag positions.

The second step involved checking quantities, such as the number of annotated documents that could be used. As we previously highlighted, there were small documents with few annotations and large documents containing hundreds of annotated entities. During the execution of our annotation methodology, we progressed quickly. Once we reached 500 annotated documents, we ran experiments and validations to assess the quality, which revealed that some annotations lacked quality. We then revisited the process to refine the annotations in certain documents until we finalized the dataset. However, we first presented a version of the same dataset that lacked quality, and our experiments presented bad outcomes. Observing it, we applied a

Table 21: Entity distribution in the final small annotated Dataset specialized on COVID-19 clinical notes.

Entity (PT-BR)	**Entity (ENG))**	**Count**
Atributo Clinicos	Clinical Attributes	6765
Procedimentos Invasivo ou Terapeutico	Invasive or Therapeutic Procedure	4047
Resultado	Result	3983
Doenca ou Sindrome	Disease or Syndrome	1819
Sinal ou Sintoma	Sign or Symptom	1345
Procedimento Diagnostico	Diagnosis procedure	707
Total of entities	-	18.666
Total of documents	-	314

Source: Created by the author.

process to qualify data with the annotators. Through the qualification process, we understood that some of our annotations were too long, more than two words, considering the context more than an accepted entity.

Our small annotated dataset contains 314 documents, including both small and large clinical notes, and is quite varied. Next, we evaluated the number of annotations by the type of named entity defined. Since we were working with unstructured data from hospitals, we expected to encounter variability in the volume and types of entities across documents, which makes it challenging to normalize and balance a small dataset like ours. Table 21 summarizes our final annotated dataset to fine-tune experiments on BERT models further.

This dataset contains annotations for the Named Entity Recognition (NER) task in clinical texts, following the BIO (Beginning, Inside, Outside) convention. The O (Outside) category is the most frequent, with 52,661 occurrences representing tokens that do not belong to any specific entity. The annotated categories include Clinical Attributes, Invasive or Therapeutic Procedures, Results, Diseases or Syndromes, Signs or Symptoms, and Diagnostic Procedures, with labels indicating each entity's beginning (B-) and continuation (I-). The distribution of labels ranges from 6,755 tokens in Clinical Attributes to 707 tokens in Diagnostic Procedures, reflecting the diversity of textual clinical notes in the real-world dataset. Comprising 314 documents, this dataset covers a small range of annotated clinical texts.

However, we pre-analyzed several documents before selecting them for the annotation process. Considering the structure and the evolution characteristics, many entries accumulated parts of previous clinical notes, so even if a patient had a total of 50 clinical notes recorded on a given day of assessment, we often selected only the last clinical note of the day to annotation process, since it would contain all previous notes plus the most recent update. Thus, it became a common scenario that there would be 50 clinical notes, but we would ultimately select only 3 to 5 clinical notes. This makes it

suitable for training once we avoid repeated text entries and makes it better for NER models to generalize and identify essential information not seen, such as diagnoses, symptoms, and medical procedures in the COVID-19 domain. Facilitating extracting relevant information from clinical notes in Portuguese, in the public or private sector, and allowing health record analysis.

For this experiment, we started with the Bert-base-multilingual-cased because it can handle multiple languages, including Portuguese, to analyze our clinical notes in a multilingual context. This model was trained in 104 languages and released uncased and cased versions (PIRES; SCHLINGER; GARRETTE, 2019). Its transformer-based architecture allows it to capture complex relationships between words in different contexts, essential for named entity recognition (NER) in medical texts.

5.2.2 Named Entity Recognition

As described, we aimed to extract relevant information from unstructured data and organize it into a healthcare standard format, such as an ontology. We selected three models to evaluate their performance on downstream tasks to achieve this, utilizing pre-trained models from different domains over our annotated dataset. Recent studies have demonstrated the effectiveness of Transformer-based Bidirectional architectures in various information extraction tasks, yielding promising results (SCHNEIDER et al., 2020; OTTER; MEDINA; KALITA, 2020; LAPARRA et al., 2021). Additionally, as noted by Tinn et al. (2021), in the health domain, datasets are often small, limiting the ability to train large models, which is also the case in our study of COVID-19 clinical notes.

To address this, we explored three BERT models to evaluate their performance on our small annotated dataset, validating the effectiveness of machine learning algorithms to extract information from COVID-19 clinical notes. As highlighted by Schneider et al. (2020), models pre-trained on domain-specific data can enhance the performance of downstream tasks. However, considering our research goal over a small annotated dataset of clinical notes in Portuguese focused on COVID-19, there is not a pre-trained model aligned to this domain in Portuguese. In this sense, we experimented with models from a general domain, as is the case of Bert-base-multilingual-cased, Bert-base-portuguese-cased as well as a pre-trained model in a specific domain, such as *BioBERTPt-Clin*.

In the one hand, the Bert-base-multilingual-cased (M-BERT)[11] and Bert-base-portuguese-cased[12], were open-domain training data. The first on the top 104 languages with the largest Wikipedia dataset, therefore, making it generalist (PIRES; SCHLINGER; GARRETTE, 2019) and, the second, as demonstrated by Souza (2020),

[11]M-BERT <https://huggingface.co/bert-base-multilingual-cased>
[12]Bert-Portuguese<https://github.com/neuralmind-ai/portuguese-bert>

pre-trained on Portuguese data, outperforms the best-published results in Portuguese experiments, but remains generalist on Portuguese.

On the other hand, in the BioBertPT-Clin specific-domain model, proposed by Schneider et al. (2020), the authors fine-tuned M-BERT using Portuguese clinical notes from the narratives of Brazilian hospitals and biomedical abstracts from Scielo[13] and PubMed[14] papers. BioBERTpt represents a domain-specific model to experiment with, and it reached the state-of-the-art on the CLINpt corpus. However, the dataset used to pre-train the model was general clinical data from an electronic health system instead of specialized COVID-19 data, which can interfere with classification quality. The model showed promising results for NER tasks in the health domain, showing good evaluations on clinical notes SemClinBr corpora, unstructured clinical notes (SCHNEIDER et al., 2020; SOUZA, 2020). Through these experiments, we evaluated our annotated dataset on three different BERT models, allowing us to understand how the differences in pre-training domains can cover and influence information recognition and classification, even by applying fine-tuning to a small dataset (NASAR; JAFFRY; MALIK, 2021; RIVERA-ZAVALA; MARTÍNEZ, 2021).

5.2.3 Evaluation metrics on Named Entity Recognition

We adopted a separate approach for each pipeline component to evaluate the prototype. We prepared and annotated the data for training on the Annotation step, and to evaluate the annotated data, we involved the annotator team. We conducted a manual evaluation with a team division to determine discrepancies in the annotated data with our specialists and students. The most practical approach requires manual assessment to ensure that the annotated data is high quality and reliable to train the experiment using ML models. In this sense, we defined a sample random splitting to perform a manual evaluation, splitting it into small sets of annotated data and dividing the team to verify the parts separately and discuss conflicts with the specialists who do not annotate.

Likewise, we used metrics frequently applied evaluations for the NER experiments, such as Precision, Recall, and F1 Score. The Named Entity Recognition experiments focused on recognition and classification tasks. We described the Classification Metrics we adopted in Table 22 since the evaluation metrics often directed to this format of experiments are related to the confusion matrix classification. Regarding these experiments, we selected classification metrics presented in Table 23, aiming to evaluate recognized entities.

Furthermore, all evaluation metrics were applied following the machine learning

[13]Scielo<https://scielo.org/>
[14]<https://pubmed.ncbi.nlm.nih.gov/>

Table 22: Confusion matrix for binary classification

	Actual positive	**Actual negative**
Predicted positive	True positive (TP)	False positive (FP)
Predicted negative	False negative (FN)	True negative (TN)

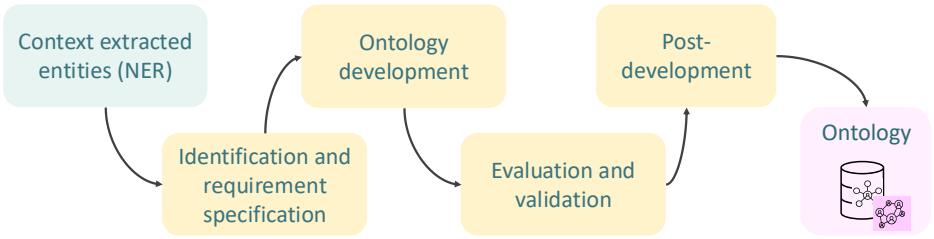
Source: Created by the author.

Table 23: Classification metrics from confusion matrices.

Measure description	**Equation**
Accuracy (ACC): the proportion of correct predictions (both true positives and true negatives) to the total number of predictions.	$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.1)$$
Precision (P): the proportion of correct positive predictions (true positive) to all the positive predictions (true positive and false positive).	$$Precision = \frac{TP}{TP + FP} \quad (5.2)$$
Recall (R): the proportion of correct positive predictions (true positive) to all the positive samples.	$$Recall = \frac{TP}{TP + FN} \quad (5.3)$$
F1 Score: the harmonic mean of precision and recall.	$$F1 = \frac{2P * R}{P + R} \quad (5.4)$$

Source: Created by the author.

Figure 20: Ontology development methodology applied to build the Covid-19 Clinical Note ontology.



Source: Created by the author.

models' approach to Data Splitting (GOODFELLOW; BENGIO; COURVILLE, 2016). We evaluated model performance using training samples (also called in-sample), which are retrodictive rather than predictive. Without testing models on future samples, there's a risk of overfitting (learning too much from the data, making too specialized) or underfitting (learning too little from the data). In our experiments, we divided the dataset into three sets: training, validation, and test sets. The training set is used to build the models, the validation set is employed to refine the models (acting as an intermediate test), and the test set is used to evaluate the final performance of the models.

5.3 Ontology structure

Our proposal to develop an ontology as the final data representation structure is based on its flexibility to harmonize with other standards (MA et al., 2023). We designed an ontology for the COVID-19 clinical domain, tailoring it to our specialized scope. This personalized approach allowed us to focus on the data we aim to represent from clinical notes while leveraging existing ontology resources. We followed the methodology proposed by Sharma and Jain (2024) to define an ontology and incorporate existing resources. The authors outlined four steps for ontology development: (1) identification and requirement specification, (2) ontology development, (3) evaluation and validation, and (4) the post-development phase, as shown in Figure 20, the complete process to define and create the ontology.

In the first step, we mapped the entities defined for the Named Entity Recognition (NER) experiments as concrete classes for the ontology structure since our goal was to represent all the data extracted from unstructured clinical notes. After this definition, we established the appropriate relationships for the existing data, reviewed relevant literature, and considered potential future alignments with established ontologies.

Considering the next step, after identifying the specifications and developing our primary structure to the ontology with classes and relationships, in phase (3), we

conducted a previous review of COVID-19 ontologies to check existent resources. This review involved consulting BioPortal[15], a comprehensive repository of biomedical ontologies (NOY et al., 2009), and conducting a non-exhaustive search of studies on Google Scholar using the snowballing methodology to find studies that published full structure of ontologies on COVID-19 domain. We applied the snowballing method to level 3, considering the recent context of COVID-19. When we finished the first phase, we found four candidate studies that published ontologies related to our context of clinical data.

We selected the four ontologies: COVID-19 Ontology, COVID-19 Ontology for Data-driven Operationalization (CODO), Coronavirus Infectious Disease Ontology (CIDO), and Covid-O Ontology developed by (SARGSYAN et al., 2020; DUTTA; DEBELLIS, 2020; HE et al., 2020; SHARMA; JAIN, 2024) respectively. Those previously published ontologies were evaluated by specialists and validated on use cases in COVID-19 and available on BioPortal. By establishing alignments with their defined concepts, we can evaluate our defined ontology classes as solid concepts, considering their established and semantically consolidated context. Finally, after the developed ontology had been finished, we used the fourth step (4) to represent all extracted data from the clinical notes.

5.3.1 Semantic matching and lexical normalization

The need for semantic matching arises from the variability encountered in recognized entities. When an entity is associated with multiple conceptual structures for the same term, it introduces variability, which is common in unstructured data. Throughout annotating the dataset and performing experiments, we identified numerous instances of terms being misspelled or misused, which is a common issue in unstructured clinical notes. Lexical normalization addresses this problem by consolidating terms based on their semantic contexts, reducing ambiguity.

We implemented a semantic matching approach to handle this variability, allowing us to standardize terms and resolve ambiguities. Medicine students who participated in the annotation process and have the expertise from contextual knowledge of COVID-19 conducted the semantic analysis. The process included generating tables with extracted entities, performing semantic analysis to align the terms, grouping terms for normalization, and creating a thesaurus in the context of COVID-19 using the Simple Knowledge Organization System (SKOS) standard. In parallel, we analyzed ontologies to evaluate potential structure alignments with existing resources representing this variability context. These alignments were compared to ensure the ontology we developed for COVID-19 clinical notes adhered to semantic consistency and met the interoperability requirements of healthcare standards. This process improved the quality

[15]BioPortal <https://bioportal.bioontology.org/>

of extracted information and provided a structured framework for future adaptations and enhancements.

Although the proposed model outlines the conceptual components needed to achieve semantic interoperability, we also explored ways to evaluate its effectiveness. To this end, we proposed developing a prototype to transform unstructured data into structured data within the healthcare domain. The first step, Dataset Design, involved preparing raw clinical note data and creating an annotated dataset with input from specialists. We conducted information extraction tasks using this annotated dataset, applying them to train transformer models to classify and extract relevant data. We employed three BERT models in Named Entity Recognition experiments to classify and extract entities. The extracted data was then organized into a Graph Structure, representing the information as an RDF ontology, which served as the final structure for the COVID-19 clinical notes. By defining an ontology as an intermediary data representation format, we ensured flexibility in selecting the most appropriate healthcare standard, allowing for future conversion and adaptation based on specific needs.

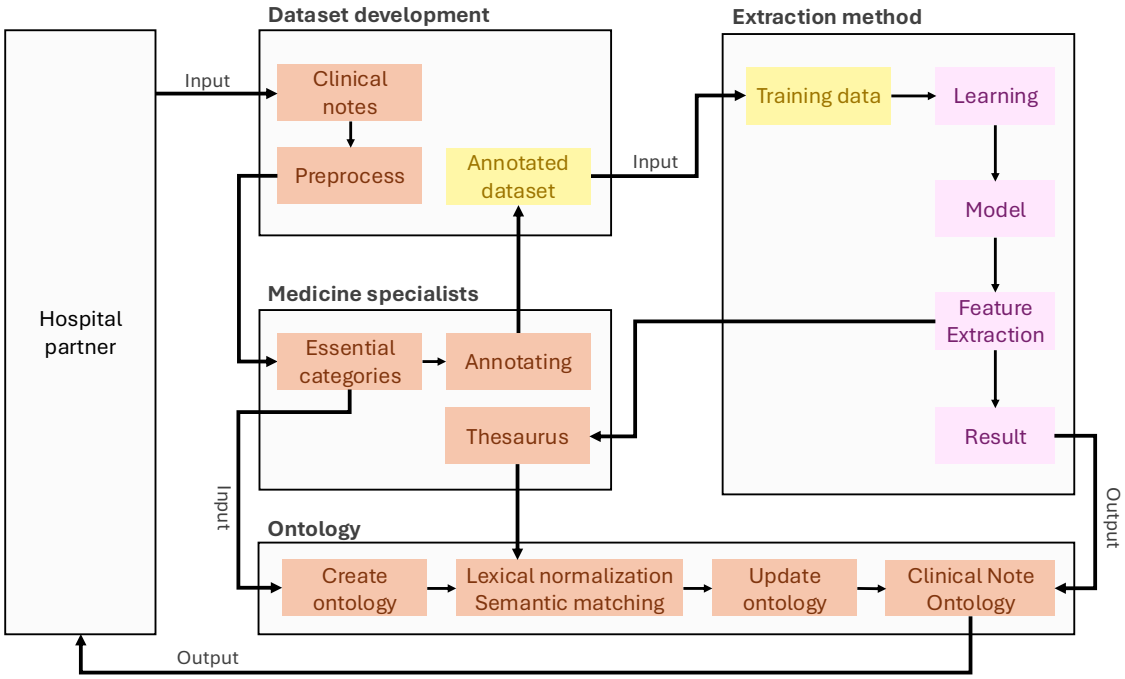
5.4 Final Remarks

As previously shown in 5.1, we executed a set of experiments to evaluate our proposed model through a prototype in a case study of COVID-19 with data provided by a hospital partner. The hospital is from Porto Alegre, a city in the South region of Brazil; more details and analysis of the dataset are given in Chapter 6. However, the pipeline proposal to evaluate the model contemplates the technological choices the model can adopt to be implemented. Nonetheless, as highlighted in Figure 21, the proposed model is independent of technological choices, as a good practice reinforced by the authors Benson and Grieve (2021), since interoperability is a concept to be implemented regardless of the selected standard, technological choices and data structure or format.

First, **Hospital partners** represents any healthcare establishments, health institutions, healthcare units, and clinics from the private or public sector that generate clinical information, with or without healthcare standard adoption within. Following the flow, considering the nature of the medical and healthcare domain, data are usually specialized, which requires a process of definition and development of a specific domain dataset to use with the information extraction methods. In this sense, the **Dataset development** is the process of annotating data to develop and provide a small and qualified dataset based on the data characteristics.

Involving specialists is a vital component of the proposed model and an essential step in achieving interoperability and healthcare standard adoption (ISO13606, 2019). The **Medicine specialists** conduct the data analysis and define the essential information that the data can provide. Then, with the developed dataset finalized, the **Extraction**

Figure 21: The complete process executed over the proposed model.



Source: Created by the author.

method applies the machine learning algorithms to extract essential information defined by specialists. The specialists' interaction also happens during the stage post-extraction method because the extracted data is analyzed to generate a specialized thesaurus, looking at lexical normalization. According to Zha et al. (2024), it highlights the potential of applying the Named Entity Normalization to standardize different representations of the same medical entity and how it substantially impacts NER performance.

Defining **Ontology** as the representation structure reinforces our commitment to technological autonomy without requiring the use of proprietary databases or relational structures, as well as reiterating the flexibility and semantic aspects that the use of an ontological structure provides (HITZLER, 2021), allowing robust representations of the complex relationships of language and health data needs. On the other hand, the benefits of using ontologies in the health area also reinforce our choice, allowing the data extracted and accommodated in the proposed ontology to be extended to other consolidated ontologies (HE et al., 2022; SHARMA; JAIN, 2024; SARGSYAN et al., 2020; HEYMANN; SHINDO, 2020). In conclusion, applying our proposed model is concerned with being extensible to the domain data, ensuring that it applies to any healthcare context. The model is also concerned with scalability and allowing integration into other representation resources, such as terminologies, taxonomies, or other clinically extracted ontologies. Therefore, the model presents aspects of flexibility as a distributed architecture that applies the components as needed.

6 RESULTS AND DISCUSSION

In this chapter, we present model evaluation results through the experiments on the COVID-19 case study, focusing on the proposed model for semantic interoperability of data in the healthcare domain. The model preprocesses unstructured clinical notes in Portuguese from EHR systems. The data provided were related to inpatients with positive COVID-19 results. Our final objective with the model is to provide an ontology structure populated with essential extracted information from the clinical notes, organized in a format easily harmonized later to healthcare standards, like openEHR, HL7 FHIR, and others. To evaluate our structure, we correlate and extend the main ontology classes to existent and consolidate ontologies in the COVID-19 context. In Section 6.1, we described the developed dataset and the clinical notes the hospital partner provided to the COVID-19 study case. In Section 6.2, we executed experiments to evaluate three versions of the developed dataset using the Bert-base-multilingual-cased model. In Section 6.3, experiment with a model pre-trained on generalized Portuguese data to evaluate performance in a different domain over our best-evaluated dataset. In Section 6.4, we experimented with a model pre-trained on the clinical domain, data from a Brazilian Portuguese hospital. Section 6.5 described the developed ontology for representing the extracted data from clinical notes. Finally, in Section 6.6, we consolidated our model outcomes and challenges to further research.

6.1 Dataset Analysis

Our curated dataset was based on data from the Conceição Hospital Group (GHC), covering the period from December 28, 2020, to December 5, 2021. It included 1401 unique patients and 65,535 clinical note entries, where several inpatients had a collection of related clinical notes in their medical records. The data collected focused on patients who tested positive for COVID-19 through RT-qPCR tests when hospitalized, providing the critical reality of unstructured data. Finally, our annotated dataset comprises 314 documents from the 65,535 entries, excluding from the 65,535 clinical notes 20,129 entries about shift change descriptions of healthcare professionals containing around 150 characters and no clinical information, representing 18,666 unique annotated entities.

GHC is a reference in the Unified Health System (SUS), which comprises multiple healthcare facilities. Operating primarily in the North Zone of Porto Alegre, four hospital units, an Emergency Care Unit[1], twelve (12) primary care units, three Psychosocial Care Centers[2], and the GHC School provide essential services. GHC ensures universal and free-of-charge access to healthcare, upholding the right to health for all. As a branch of

[1]in Portuguese, Unidade de Pronto Atendimento UPA
[2]in Portuguese, Centros de Atenção Psicossocial CAPS

Table 24: Covid-19 statistics across different regions in Brazil.

Region	Population	Accumulated cases	Accumulated deaths
Southeast	88.371.433	8.662.788	294.659
Northeast	57.071.654	4.950.096	120.019
South	29.975.984	4.348.954	97.519
North	18.430.980	1.923.911	47.548
Central-West	16.297.074	2.401.772	59.311
Total	210.147.125	22,287.521	618.056

Source: Created by the author based on 2021 Covid-19 statistics in Brazil[a].

[a]<https://infoms.saude.gov.br/extensions/covid-19_html/covid-19_html.html>

the Ministry of Health, GHC is the largest public hospital network in southern Brazil, providing 100% SUS-based care. With 1,391 beds, it manages 46,100 hospitalizations annually for Rio Grande do Sul (RS) residents, showing its operations' scale and impact. GHC's team of 9,874 professionals delivers approximately 1.1 million consultations and 27,000 surgeries annually. The group conducts around 4 million exams annually and is responsible for diagnosing over half of the cancer cases expected in the Porto Alegre population.

The clinical notes data provided did not include explicit information regarding the patient's demographic details, such as their place of residence or whether they originated from other cities or states. As a result, we cannot definitively determine the geographical origin of the hospitalized patient. However, it is essential to highlight that the structure of the clinical notes reflects a regional healthcare scenario specific to the state of Rio Grande do Sul (RS), thus giving a localized context to the data in Porto Alegre. The regional scenario of our data, provided by a reference hospital in the RS capital, is a characteristic of stratification for data. This represents a research limitation in applying the developed dataset broadly, and it might need improvements to be largely adopted by other public hospitals in the context of Brazil.

According to the Brazil Ministry of Health COVID-19 monitoring panel[3] as shown in Table 24, Brazil 2021 scenario accumulated 22,287,521 cases while 619,056 accumulated deaths. Considering our regional data scenario, RS is one of the three states representing the South region and has an estimated population of 11,377,239 in 2021. By the same year, the state reported 1,507,117 accumulated cases of COVID-19, alongside 36,444 accumulated deaths. These data provide a critical context for our dataset, reflecting the substantial impact of the pandemic in the region. While the clinical notes data may not specify patient origins, this broader regional data frames the significance of the pandemic's burden on the healthcare system in RS and gives a sense of the scale of hospitalizations and medical interventions during that time.

[3]Ministry of Health COVID-19, accessed on October 1, 2024<https://infoms.saude.gov.br/extensions/covid-19_html/covid-19_html.html>

Table 25: Experiment performed using Bert-base-Multilingual on the first version of our dataset.

Entity	**Precision**	**Recall**	**F1-score**	**Support**
Clinical attributes	0.27	0.41	0.32	63
Disease or syndrome	0.55	0.70	0.62	102
Diagnosis procedure	0.33	0.16	0.21	19
Invasive or therapeutic procedure	0.30	0.49	0.37	174
Result	0.00	0.00	0.00	27
Sign or symptom	0.28	0.36	0.31	36
Micro Average	0.36	0.46	0.41	-
Macro Average	0.29	0.35	0.31	-
Weighted Average	0.35	0.46	0.39	-
Total of documents	-	-	-	252

Source: Created by the author based on the experiment classification report.

6.2 Named Entity Recognition using Bert-base-Multilingual-Cased

We used consistent training parameters across all experiments to ensure comparable results. The maximum sequence length (MAX_LEN) was set to 128 tokens, and the model was trained with a batch size of 8 for training (TRAIN_BATCH_SIZE) and 8 for validation (VALID_BATCH_SIZE), a careful choice likely made to manage GPU memory limitations. The training spanned five epochs (EPOCHS), providing the model ample time to learn patterns in the data. A learning rate (LEARNING_RATE) of 3e-05, commonly used in BERT-based models, ensured slow and precise weight adjustments. Additionally, gradient clipping was set to a maximum of one (1) (MAX_GRAD_NORM) to prevent exploding gradients and improve model stability.

6.2.1 Experiment 1 using Dataset version 1.0 from 2023

As our first experiment, we ran the BERT Multilingual Cased model on the first version of our dataset to check its quality. Although it was very small, which made any relevant extractions impossible, we could extract information from the dataset to track notes issues. Table 25 shows the distribution of entities and evaluation metrics, such as precision, recall, and F1-score, for this first version of the experiment.

Our annotated dataset shows mixed performance across the different categories of clinical entities. For **Clinical attributes**, a precision of 27% indicates that the model is making many incorrect predictions in this class. In comparison, a recall of 41% suggests it identifies a reasonable number of instances. The F1-score of 0.32 reflects the low precision. At the same time, **Disease or syndrome**, the model shows more robust performance, with a precision of 55% and a recall of 70%, achieving an F1-score of 0.62, the best among the entities, indicating that the model understands these entities well.

Table 26: Experiment performed using Bert-base-Multilingual on the second version of our dataset.

Entity	**Precision**	**Recall**	**F1-score**	**Support**
Clinical Atributtes-1	0.63	0.49	0.55	160
Disease or syndrome-1	0.42	**0.85**	0.56	71
Disease or syndrome-2	**0.88**	0.82	**0.85**	226
Diagnosis procedure-1	**0.36**	**0.33**	**0.35**	12
Diagnosis procedure-2	0.20	0.10	0.13	21
Invasive or therapeutic procedure-1	0.37	**0.52**	**0.43**	27
Invasive or therapeutic procedure-2	0.70	0.66	0.68	159
Result	0.57	0.47	0.51	88
Sign and symptom	**0.90**	0.38	**0.54**	68
Micro Average	**0.65**	**0.62**	**0.63**	-
Macro Average	0.50	0.46	0.46	-
Weighted Average	**0.68**	0.62	**0.63**	-
Total of documents	-	-	-	419

Source: Created by the author based on the experiment classification report.

The results for **Diagnosis procedure** and **Invasive or therapeutic procedure** reveal significant challenges. For **Diagnosis procedure**, the model has only 33% precision and a very low recall (16%), resulting in an F1-score of 0.21, suggesting issues in correctly identifying these entities. The **Invasive or therapeutic procedure** entity shows a precision of 30% but a higher recall of 49%, indicating that while the model identified many entities, it made numerous incorrect predictions. For Outcomes, precision, recall, and F1-score, all at 0%, suggesting the model failed to correctly identify this entity, possibly due to insufficient examples or difficulty distinguishing these entities in the text.

The overall performance of the model, represented by a micro-average F1-score of 0.41, indicates a moderate level of success in the NER task but with large variations between classes. The weighted average accuracy of 35% and F1-score of 0.39 suggest that the model lacks generalization across multiple classes, particularly in categories such as **Result** and **Sign or symptom**. The macro metrics, with an F1-score of 0.31, reinforce that underrepresented or more difficult classes, such as **Result** and **Diagnosis procedure**, are dragging down the overall performance. This indicates the need to fine-tune the model with more examples for these entities or to refine the training process to improve overall accuracy.

6.2.2 Experiment 2 using Dataset version 2.0 from 2023

In the second experiment, as shown in Table 26, we have more mixed results in terms of performance. The **Clinical Atributtes-1** entity achieved a precision of 63%, indicating

a considerable improvement in the model's ability to identify entities in this category correctly. However, the recall of 49% reveals some difficulty in identifying all instances of this entity. The **Disease or syndrome-1** entities also showed positive results, with the **Disease or syndrome-2** variant achieving an F1-score of 0.85, reflecting a good balance between precision and recall. In contrast, the general **Disease or syndrome-1** entity achieved a moderate performance with an F1-score of 0.56. There are two similar entities to **Disease or syndrome-1** and **Disease or syndrome-2**; it was a noise during the annotation process, which we had to evaluate with our specialists later to correct to release the final version of the dataset. The same issue occurred on the **Invasive or therapeutic procedure-1** and **Invasive or therapeutic procedure-2**.

On the other hand, the **Diagnosis procedure-2** or **Invasive or therapeutic procedure-2** categories continue to present challenges, with F1-scores ranging from 0.13 to 0.68. The **Result** entity performed moderately, with an F1-score of 0.51, but with more balanced precision and recall compared to the first experiment. Overall, the weighted average for the model showed an improvement in the F1-score, reaching 0.63, reflecting better performance in some entities, although others still require adjustments and more training examples.

Compared to the first experiment, there was a significant improvement in the main entities, such as **Clinical Atributtes-1** and **Disease or syndrome-2**, which saw considerable increases in F1-scores. The model's overall performance improved, as reflected in the weighted average F1-score of 0.63, compared to 0.39 in the first experiment. Despite these improvements, certain entities, such as **Diagnosis procedure-2**, continue to have low F1 scores, indicating that the model still faces difficulties in these entities and requires further refinements, possibly with more data or adjustments in the training process.

6.2.3 Experiment 3 using the final version of our Dataset from 2024

In the third experiment, as shown in Table 27, the model achieved more robust results, particularly in critical categories such as **Disease or syndrome**, which showed a precision of 82%, recall of 91%, and an F1-score of 0.87. This high performance reflects the model's ability to correctly identify most instances of this entity, especially after validating the annotations with experts. The **Clinical Attributes** category also showed notable improvement, with an F1-score of 0.64, suggesting that the validation process helped refine the annotations, contributing to more accurate recognition of these entities.

However, the **Result** and **Sign or Symptom** categories continue to pose challenges, with F1-scores of 0.47 and 0.43, respectively. The particularly low recall for Result 37% indicates that the model still has difficulties recognizing all occurrences of this entity,

Table 27: Experiment performed using Porg-base-Multilingual on the final version of our dataset.

Entity	Precision	Recall	F1-score	Support
Clinical attributes	0.60	0.69	0.64	152
Disease or syndrome	**0.82**	**0.91**	**0.87**	173
Diagnosis procedure	0.69	0.52	0.59	21
Invasive or therapeutic procedure	**0.83**	**0.84**	**0.84**	109
Result	0.62	0.37	0.47	54
Sign or symptom	0.48	0.38	0.43	26
Micro Average	**0.72**	**0.74**	**0.73**	-
Macro Average	0.67	0.62	0.64	-
Weighted Average	**0.72**	**0.74**	**0.72**	-
Total of documents	-	-	-	314

Source: Created by the author based on the experiment classification report.

possibly due to the complexity of these annotations. While validation with experts improved the quality of the annotations, additional refinement of the model with more examples may be necessary for these classes.

Compared to the previous experiments, the third experiment shows a significant improvement, particularly in terms of F1-score and weighted average, which increased to 0.72. Validation with experts contributed to a substantial enhancement in the quality of the annotations, especially in the most frequent classes, such as **Disease or syndrome** and **Clinical Attributes**, while still highlighting the need for further adjustments in underrepresented entities, such as **Result** and **Sign or symptom**.

Overall, as Table 28, the third experiment presented the best results in almost all categories, with a particularly strong performance in **Disease or syndrome** and **Invasive or therapeutic procedure**, which achieved F1-scores of 0.87 and 0.84, respectively. These scores significantly surpass those obtained in previous experiments. The validation of annotations by experts and the refinement of the dataset, even with fewer documents in the third experiment (314 versus 419 in the second), clearly contributed to a substantial improvement in the model's performance. The careful curation of the dataset ensured that the model was trained with more accurate and representative data, resulting in better generalization and greater confidence in the predictions. This emphasizes the importance of prioritizing the quality of annotations and data over quantity, demonstrating that a smaller but more accurately annotated dataset can lead to significantly superior performance.

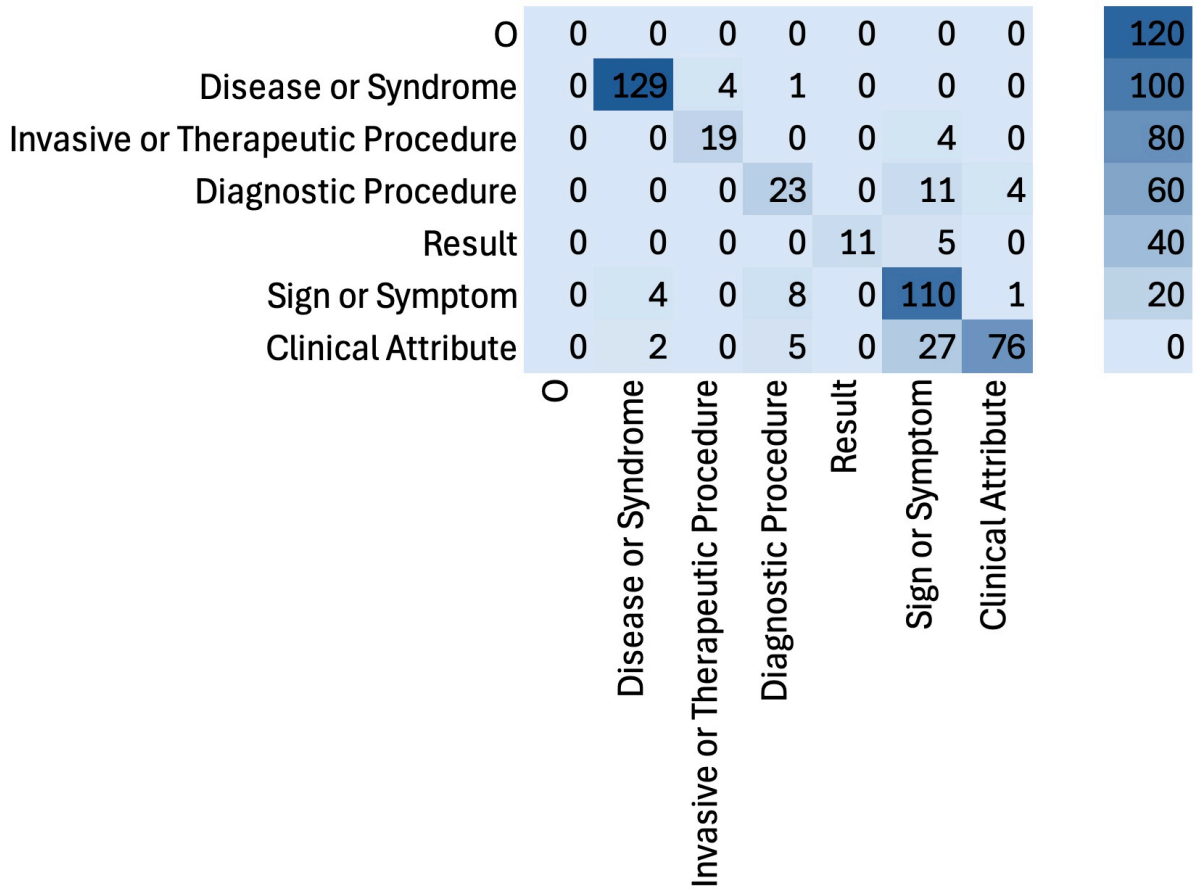
We executed a set of experiments over the three (3) versions of our dataset, showing that our validated and qualified dataset in the last version achieved better results, as shown in Table 28, **Exp3 F1** column. Conversely, as demonstrated by the Confusion Matrix in Figure 22, the first two experiments did not perform as well. In this scenario, we also tested our dataset on two additional models: Bert-base-Portuguese-Cased and the

Table 28: Comparison of classification results across the three experiments performed applied on Bert-base-Multilingual.

Entity	Exp1 F1	Exp2 F1	Exp3 F1	Exp1 Support	Exp2 Support	Exp3 Support
Clinical attributes	0.32	0.55	**0.64**	63	160	152
Disease or syndrome	0.56	0.56	**0.87**	71	71	173
Diagnostic procedure	0.21	0.35	**0.59**	19	12	21
Invasive or therapeutic procedure	0.37	0.68	**0.84**	74	159	109
Result	0.00	**0.51**	0.47	27	88	54
Sign or symptom	0.31	**0.54**	0.43	36	68	26
Micro avg	0.41	0.63	**0.73**	321	834	535
Macro avg	0.31	0.46	**0.64**	321	834	535
Weighted avg	0.39	0.63	**0.72**	321	834	535
Total of Documents	252	419	314	-	-	-

Source: Created by the author based on the classification reports.

Figure 22: Confusion Matrix generated over the experiment using Bert-base-Multilingual on the final version of our Dataset.



Source: Created by the author based on the experiment classification report .

Table 29: Experiment performed using the final version of our dataset to fine-tune the Bert-base-Portuguese-Cased.

Entity	**Precision**	**Recall**	**F1-score**	**Support**
Clinical attributes	0.61	**0.79**	0.69	126
Disease or syndrome	**0.84**	**0.92**	**0.88**	135
Diagnostic procedure	**0.78**	0.41	0.54	17
Invasive or therapeutic procedure	**0.89**	0.80	**0.84**	80
Result	0.67	0.46	0.55	39
Sign or symptom	0.62	**0.71**	0.67	21
Micro Average	0.74	**0.78**	0.76	418
Macro Average	0.73	0.68	0.69	418
Weighted Average	**0.75**	**0.78**	0.76	418
Total of documents	-	-	-	314

Source: Created by the author based on the classification report for BERT-base-Portuguese-Cased.

Brazilian model trained on clinical databases, BioBERTpt-Clin, to evaluate the dataset and to verify the different performance in domain-specific trained transformers models.

6.3 Named Entity Recognition using Bert-Base-Portuguese-Cased

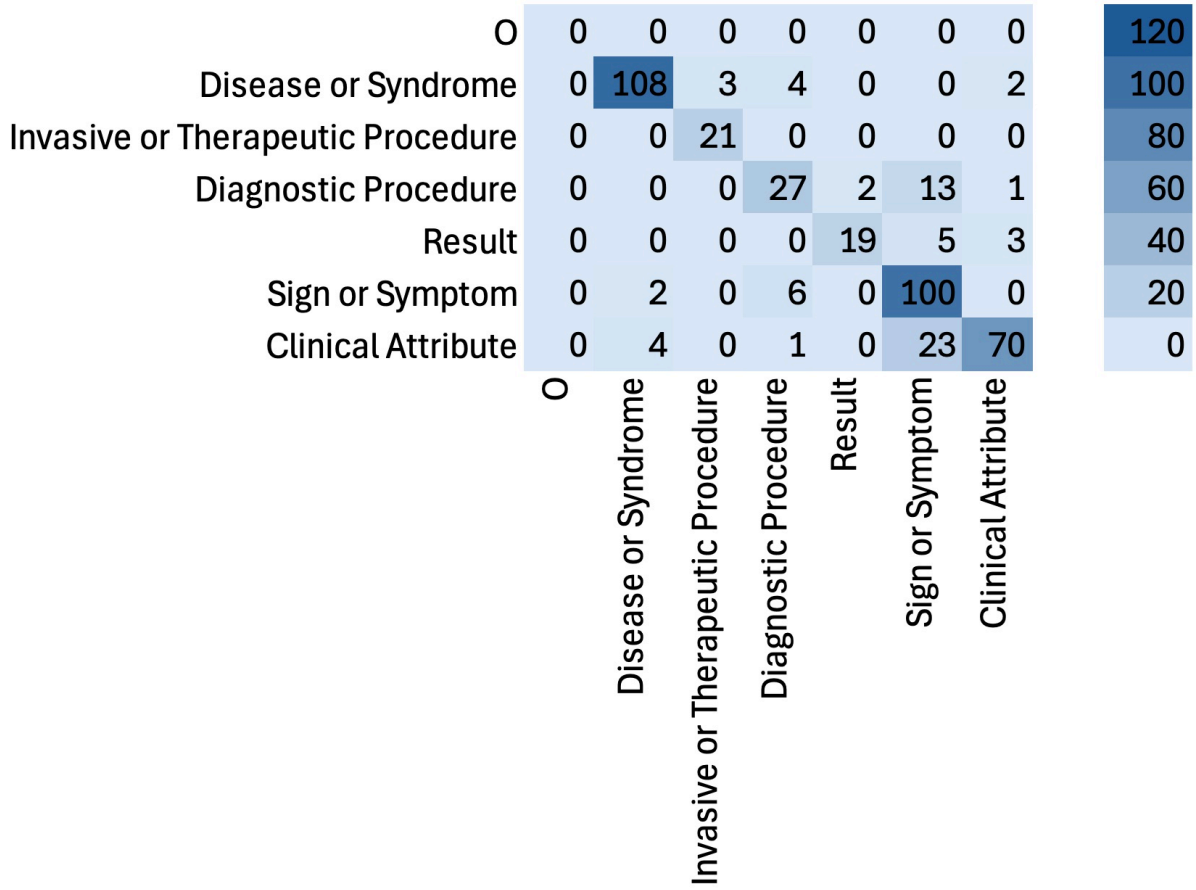
After evaluating the dataset with the Bert-base-multilingual-cased model, we executed a set of other experiments on the dataset's last version to verify performance in different pre-training domain models.

We experimented with the Bert-base-Portuguese-Cased; the overall performance with the model was quite positive, especially in the main classes. Table 29 shows how the model performed, and we can see **Disease or syndrome** entity performed the best, with a precision of 84%, recall of 92%, and an F1-score of 0.88, demonstrating that the model was highly effective in correctly identifying these entities. Likewise, the **Invasive or therapeutic procedure** class also performed well, with an F1-score of 0.84, confirming the model's ability to identify most instances of this entity.

On the other hand, some classes faced challenges; as we can see, the **Diagnostic procedure** category had a precision of 78%, but the recall was only 41%, resulting in an F1-score of 0.54. This suggests that, although the model made correct predictions for most examples classified as **Diagnostic procedure**, it still failed to identify a significant portion of these instances. The **Result** entity also showed average performance, with an F1-score of 0.55, indicating that the model still struggles to identify these entities.

Overall, the weighted averages reflect stable performance, with a Micro Average of 0.76 and a Weighted Average of 0.76, indicating a good balance between precision and recall across the dataset. Using Bert-base-Portuguese-Case has proven to be an effective choice for processing clinical texts in Portuguese, particularly in the categories with the highest support. However, there is still room for improvement in smaller and more

Figure 23: Confusion Matrix generated over the experiment using Bert-base-Portuguese-Cased.



Source: Created by the author based on the experiment classification report .

complex classes, such as **Diagnostic procedure** and **Result**, to achieve more consistent performance across all entities.

By analyzing the confusion matrix in Figure 23 from the experiment, we observed good precision in the predictions for the classes with greater support, such as **Disease or syndrome** , where correct results (main diagonal) dominate, with only a few minor misclassifications. With many correct predictions, the **Clinical Attributes** class also stands out positively. However, there are issues between the **Sign or symptom** class, where prediction errors are more common. This indicates that the model still has difficulties distinguishing some categories, especially in Clinical attributes and Result, where incorrect classifications are more scattered. Overall, the model performs well in classes with greater support but faces challenges in classes with lower support or more confusing categories.

Table 30: Experiment performed using the BioBERTpt-Clin model.

Entity	Precision	Recall	F1-score	Support
Clinical attributes	0.69	0.75	0.72	133
Disease or syndrome	**0.89**	**0.97**	**0.92**	145
Diagnostic procedure	**0.89**	**0.84**	**0.86**	19
Invasive or therapeutic procedure	**0.90**	**0.79**	**0.84**	**82**
Result	0.68	0.55	0.61	38
Sign or symptom	0.57	0.64	0.60	25
Micro Average	0.79	0.81	0.80	442
Macro Average	0.77	0.76	0.76	442
Weighted Average	0.80	0.81	0.80	442
Total of documents	-	-	-	314

Source: Created by the author based on the classification report for BioBERTpt-Clin and the pre-trained model of Schneider et al. (2020).

6.4 Named Entity Recognition using BioBERTpt-Clin

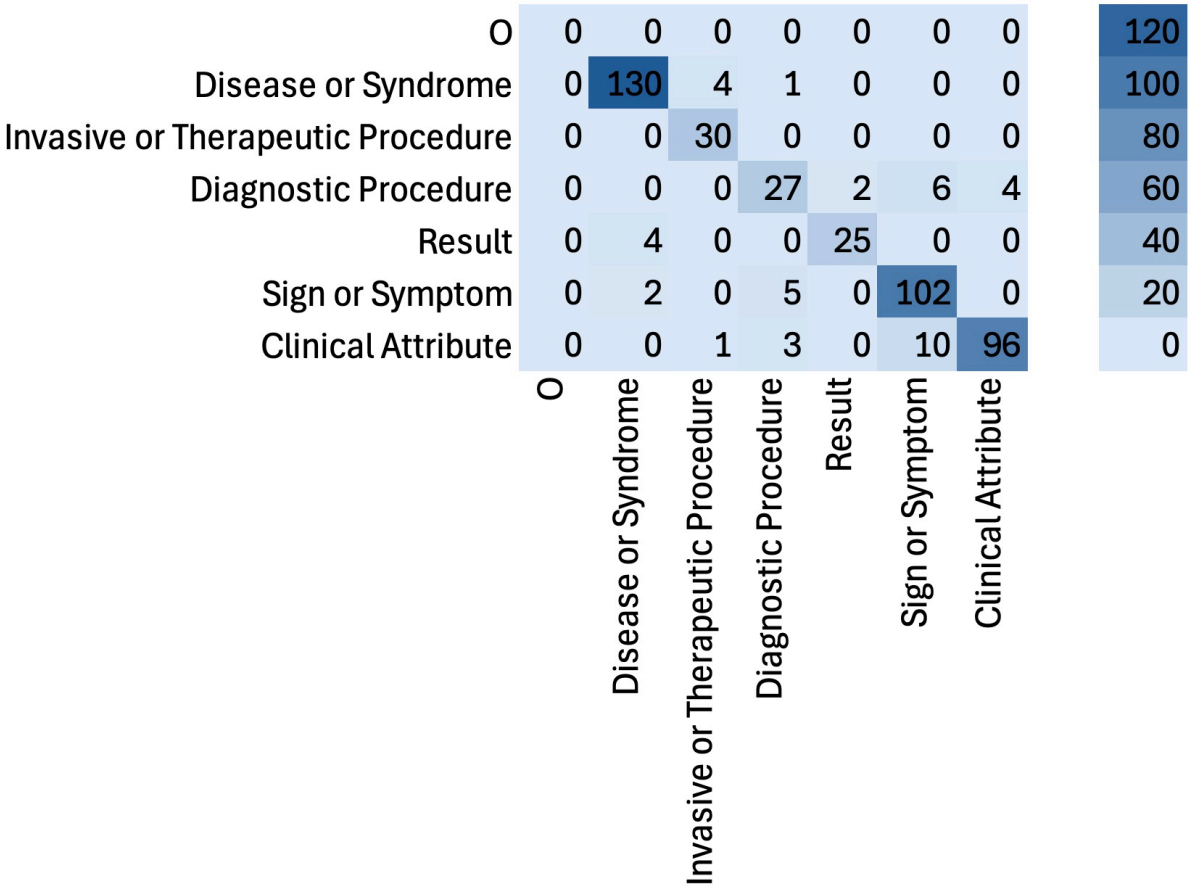
The experiment using the BioBERTpt-Clin model, specifically trained in the healthcare domain, produced promising results. As presented in Table 30, the entities related to **Disease or syndrome** performed best, with the **Disease or syndrome** entity achieving an F1-score of 0.92 and the P**Diagnostic procedure** reaching 0.86. The model also demonstrated considerable precision for **Invasive or therapeutic procedure** with an F1-score of 0.84, highlighting its ability to recognize technical terms specific to the medical field.

The entities **Clinical attributes** and **Result** showed reasonable performance, with F1-scores of 0.72 and 0.61, respectively. The slightly lower performance for **Sign or symptom** with an F1-score of 0.60, indicates room for improvement in recognizing clinical signs and symptoms, areas involving ambiguous terms, and varied contexts. However, the model maintained a robust weighted average performance of 0.80, demonstrating its overall effectiveness in classifying clinical entities.

Compared to previous experiments, BioBERTpt-Clin substantially improves its identification of health-related entities. This improvement is especially notable because the model was pre-trained on healthcare-specific data, reflecting its superior performance in more technical and specialized entities. The model's familiarity with clinical terminology enabled increased precision and recall, surpassing the results from more general models like Bert-base-Multilingual-Cased, which were not optimized for clinical data.

As we can highlight from the Figure 24, the confusion matrix for the BioBERTpt-Clin model reveals a better performance in identifying clinical entities. We can observe that the model accurately identified **Clinical attributes**, with significant correct predictions. The most relevant issues occurred in entities related to **Result** and **Invasive or therapeutic**

Figure 24: Confusion Matrix generated over the experiment using BioBertpt-Clin.



Source: Created by the author based on the experiment classification report .

procedure, where the model misclassified some entities with other related categories. However, it maintained good accuracy for classes with a higher volume of data, such as **Sign or symptom** and **Clinical attributes**.

Applying a fine-tuned BioBERTpt-Clin in the healthcare domain provided an advantage compared to other models that were not trained on clinical data. The model's familiarity with medical terminologies allows for greater accuracy, especially in complex entities like **Disease or syndrome** and **Diagnostic procedure**. Compared to previous models like BERT-base-Multilingual or BERT-base-Portuguese-Cased, BioBERTpt-Clin outperforms in terms of F1-score, demonstrating that pre-trained models in a specific domain offer more qualified information extraction and a greater ability to handle the nuances of clinical vocabulary, resulting in a confusion matrix with fewer classification errors.

6.4.1 Comparison with related work

Looking at the impact of the pandemic on clinical data generation and structuring. During the COVID-19 pandemic, the rapid emergence of new information and the need for quick decision-making posed significant challenges to handling clinical data. Raza and Schwartz (2023) observed that the massive volume of collected data—primarily unstructured—was crucial for identifying new treatments and protocols. The demand for relevant clinical information in this context underscored the importance of effective information extraction techniques, especially for systematically documenting positive cases and ensuring the clinical utility of the data.

These challenges highlight the need for more specialized healthcare models, such as BioBERTpt, and our finetuned approach focused on the COVID-19 context, which has proven more effective at handling specific medical terms and concepts. With consistent F1-scores of 0.92 for Disease or Syndrome and 0.86 for Diagnostic Procedure, models pre-trained on large medical corpora, like BioBERTpt-Clin, have been highly effective in capturing the nuances of clinical data, as supported by recent studies on clinical text information extraction (RAZA; SCHWARTZ, 2023).

On the other hand, entity extraction from unstructured clinical notes is primarily approached through rule-based systems, supervised machine learning, and hybrid techniques. Traditional rule-based methods like MedLEE, MetaMap, and cTAKES rely heavily on medical experts to manually create rules using medical knowledge bases and dictionaries. However, Yang et al. (2020) points out that the rigidity of these methods limits their adaptability to new terms and concepts, particularly in dynamic situations like the COVID-19 pandemic.

In contrast, more modern machine learning approaches, such as CLAMP, treat clinical concept extraction as a Named Entity Recognition (NER) task. Sim et al. (2023)

highlights that these approaches perform better, especially when combining deep neural networks with large language models pre-trained on clinical corpora, as seen with models like BioBERTpt. In our experiments, BioBERTpt-Clin outperformed generic models such as BERT-base Multilingual and BERT-base Portuguese-Cased, demonstrating the advantage of domain specialization for extracting complex clinical entities.

Our experiments using BioBERTpt-Clin, BERT-base Multilingual, and BERT-base Portuguese-Cased underscore the importance of specialized models for clinical information extraction. BioBERTpt-Clin, pre-trained on medical corpora and finetuned on our specialized covid-19 dataset, achieved the best overall performance with F1-scores of 0.92 for Disease or Syndrome, 0.86 for Diagnostic Procedure, and 0.84 for Invasive or Therapeutic Procedure. These results reflect a suitable generalization of models trained on healthcare data in managing medical terminology and concepts, which are often ambiguous or highly specialized. In contrast, while BERT-base Multilingual showed competitive performance with an F1-score of 0.87 for Disease or Syndrome, it struggled with more specialized entities like Diagnostic Procedures and Results, likely due to its lack of medical domain-specific training.

Kolukula et al. (2024), Zha et al. (2024), Jiang et al. (2021) argue that applying models without sufficient domain specialization can lead to confused classifications and lower precision for clinical terms. Similarly, BERT-base Portuguese-Cased, though tailored for Portuguese, showed intermediate performance with an F1-score of 0.88 for Disease or Syndrome. The lack of clinical data training affected its results, particularly with more complex entities, highlighting the need for domain-specific training to improve entity recognition, as noted by Pomares-Quimbaya, Kreuzthaler and Schulz (2019).

The impact of data balance and annotation quality was an important observation from the experiments. Models trained on limited or unbalanced datasets can show performance variability across different entities. However, the analysis revealed that Diagnostic Procedures exhibited less variation and fewer lexical errors, suggesting that even with a reduced volume of annotations, consistency and clarity in medical descriptions can ensure robust performance, as observed by Kolukula et al. (2024).

Concerning the correlation between annotations and model performance, another finding is the correlation between the number of annotated examples and model performance. For instance, the entity clinical attributes achieved a high score, directly linked to the number of cases annotated in the dataset, as presented in Table 21. Raza and Schwartz (2023) also identified that increasing the volume of annotated data improves the accuracy of information extraction models, especially in clinical contexts where data variability can be challenging. This result underscores the importance of expanding annotated data to enhance model accuracy and capacity to handle complex categories.

A criação e manutenção de um dataset especializado é onerosa e costume ser a part

do processo que mais irá consumir tepmo da equipe de especialista. Neste cenário, os autores Zha et al. (2024) tem trat

6.5 Developed Ontology

To create the ontology, we followed, as defined, the four steps guided by the authors Sharma and Jain (2024), (1) identification and requirement specification, (2) ontology development, (3) evaluation and validation, and (4) the post-development phase. Based on that methodology, we first defined a conceptual plan and specification of the ontology, followed by the development of the Protegé Tool. Next, to evaluate and ensure the solid structure defined, we filtered ontologies focused on the COVID-19 context and sought alignment with our defined classes and relationships. Finally, we delivered the ontology structured and executed experiments applying the experiment's fine-tuned BERT models to extract information from the clinical notes available and accommodate data into the ontology to a last verification and evaluation of structure.

A vital aspect considered for the development of the ontology was the context in which it would be applied to represent clinical notes information, with dynamic data and significant variability and temporal characteristics. Clinical notes are, in essence, an observation of the patient's state described by the healthcare professional who treats him/her. Beale, Heard et al. (2007) reinforce that 'observations' as any clinical information obtained by the investigator and used to characterize the patient state is 'in the past' since it expresses states or events that have already occurred. Observations are, therefore, historical in nature. In this sense, we considered everything that characterizes the patient, including signs, symptoms, procedures performed, and even test results, as temporal data related to the patient's past.

This characteristic reinforces adopting an approach employed by the main interoperability standards, such as OpenEHR and HL7 FHIR. The classes are the main structure and organization aspects of the information accommodated in the ontology, while our instances accommodate the extracted data. This approach differs from approaches that represent dynamic values as subclasses, intending to build a taxonomy or a hierarchical structure of the information accommodated in the ontology.

Looking at the OpenEHR modeling aspects, instances represent mutable data, such as the instances of archetypes. Each instance represents a sign, symptom, or diagnosis, for example. The HL7 FHIR standard also exemplifies using instances to capture dynamic clinical data. In FHIR, entities such as Patient[4], Condition[5], Procedure[6], and Observation[7] are represented as resource instances. Each instance can capture specific

[4]<https://build.fhir.org/patient.html>
[5]<https://build.fhir.org/condition.html>
[6]<https://build.fhir.org/procedure.html>
[7]<https://build.fhir.org/observation.html>

data, facilitating interoperability and management of dynamic data. The FHIR resource model treats symptoms, diagnoses, and procedures as instances that can be created, modified, and archived according to the dynamic needs of the clinical record, reflecting the evolution of the patient's health status.

As defined, in the first phase (1), we prioritized including all the information that would be recognized and extracted from clinical notes through Named Entity Recognition experiments as specifications for the ontology. In this way, the initially planned ontology concepts must accommodate our data related to the previously defined entities: clinical attribute, invasive or therapeutic procedure, result, disease or syndrome, sign or symptoms, and diagnostic procedure. We follow the same approach standards as openEHR and HL7 FHIR and accommodate our mutable information as instances instead of creating sub-classes.

In this way, we represented the initial structure of the ontology, starting from a definition of the classes as the entities extracted in the experiments. With this approach to representing data, we ensure that extracted classes and their meanings and contexts are immutable, such as a symptom independent of what kind of symptom will always be a symptom. However, dyspnea (an instance of a symptom) and shortness of breath (an instance of a symptom) are the same, and representing them as a subclass would be excellent for creating a taxonomy of symptoms. Still, our primary goal in creating the ontology is to accommodate a semantic and interoperable way for clinical note data, keeping the temporal nature and allowing the sharing of essential information about COVID-19 inpatients.

When defining the structure, it is essential to consider that the Patient is also a concrete class; even though it was not an entity directly extracted from the clinical data, all the information is directly related to this class, and thus, we maintain the semantic relationship. The same decision was made when we defined clinicalNote as a Class since all the data is accommodated as clinicalNote and related to all classes through it.

- ClinicalAttribute
- DiseaseSyndrome
- InvasiveTherapeuticProcedure
- Result
- SignSymptom
- DiagnosticProcedure
- Patient
- ClinicalNote

The defined entities for this research were based on the previous study of Oliveira et al. (2020), which annotated medical notes and delivered a dataset in Portuguese on the general clinical domain. As well as the authors, we conducted an exploratory analysis of our data to understand the main semantic context we could extract and essential information that, once standardized for interoperability, could improve continuity of care. Building on the insights from this exploratory analysis, we defined the relationships between the extracted entities and the corresponding classes, linking them semantically to ensure meaningful representation.

The Clinical Attribute class refers to specific information about a patient's physical characteristics, such as age, height, weight, or blood pressure. These attributes provide essential baseline data for understanding the patient's health. To represent this relationship, we created the semantic link: `Patient -> Clinical Attribute -> hasAttribute`, where the patient is associated with a particular attribute.

On the other hand, the Disease or Syndrome class refers to a condition that reflects a problem, a pre-existing diagnosis, or a newly identified clinical issue that requires medical attention. Following the HL7 FHIR Condition resource specification, we define this class to encompass any event, situation, or clinical concept that has reached a level that impacts the patient's care or well-being. To represent this, we established the semantic relationship: `Patient -> Disease or Syndrome -> diagnosedWith`, indicating that a patient has been diagnosed with a specific disease or syndrome.

The Invasive or Therapeutic Procedure class encompasses procedures that physically intervene in the patient's body (invasive procedures) and treatments involving the administration of medication (therapeutic procedures). These procedures can be current or historical and are performed for the patient's benefit, either directly by healthcare practitioners or through medical devices.

Invasive procedures refer to interventions that involve entering the body, such as intubation, drains or tubes, and tracheostomy. These procedures are critical in managing severe cases where airway management or fluid drainage is required. On the other hand, therapeutic procedures are associated with administering medications, such as antimicrobials, to treat infections or manage conditions. We created the semantic relationship `Patient -> Invasive or Therapeutic Procedure -> receivesMedicationTreatment` to represent these procedures within the ontology. This relationship captures the administration of therapeutic treatments, such as medications, and the execution of invasive procedures on the patient.

The Result class is related to any outcomes or data derived from laboratory tests, imaging, or other diagnostic evaluations performed on the patient. These results play a critical role in monitoring the patient's health status, tracking the progress of treatments, and identifying potential complications. The relationship `Patient -> Result -> monitorsPatient` captures the semantic link between the patient and their

test results, highlighting how these results are used to assess the patient's condition over time.

The Sign or Symptom class refers to clinical observations associated with the patient, which can be either objective signs (such as a fever detected by a thermometer) or subjective symptoms (such as fatigue reported by the patient). These clinical manifestations are essential for diagnosing the patient's condition and determining the appropriate follow-up care. The *hasSymptom* relationship, represented as `Patient -> Sign or Symptom -> hasSymptom`, links the patient to the signs or symptoms they exhibit, forming a crucial part of the clinical picture.

The Diagnostic Procedure class encompasses any examination or test to evaluate the patient's condition. This includes various procedures, such as laboratory tests (e.g.,GeneExpert, PCR, RT-PCR, QPCR, HMC, ATQ, HTG), and other diagnostic methods. These procedures involve specific devices, test types, and methods, though they do not necessarily include the results. The relationship `Patient -> Diagnostic Procedure -> performsTest` captures the act of conducting these diagnostic tests on the patient, allowing for a clear representation of the processes involved in diagnosing and managing the patient's health.

The final class, ClinicalNote, plays a critical role in our ontology as it consolidates and connects all extracted information related to the patient. Clinical notes contain essential data, such as signs and symptoms, diseases, diagnostic procedures, and treatment decisions, all documented by healthcare professionals during patient care. According to the HL7 FHIR standards, clinical notes represent a wide variety of documents generated on behalf of the patient throughout different stages of care, ensuring comprehensive coverage of the patient's medical history.

In our ontology, the ClinicalNote class links all the extracted information to the patient, providing a complete, unified view of the patient's clinical data. This class enables the representation of temporal records, which is crucial for understanding the evolution of a patient's condition over time. The temporal variable is incorporated as an annotation property, DataDeRegistro, which captures the timestamp of each clinical entry. This annotation allows for detailed tracking of when specific symptoms were observed, diagnoses were made, or treatments were administered.

The relationship `Patient -> ClinicalNote -> hasClinicalNote` semantically links the patient to their clinical notes, facilitating inference and temporal queries. By consolidating all relevant clinical information and associating it with the time of entry, this class allows the ontology to support advanced queries and analysis, such as tracing the progression of a disease or the effectiveness of treatment over time. In Table 31, we summarize the relationships and related classes.

After finalizing the definition of classes and relationships, ensuring a structured and semantically rich model that captures the dynamic nature of clinical data. Our next step

Table 31: Relationship created to semantically link classes to data extracted.

Class	**Relationship PT-Br**	**Relationship ENG**
Clinical Attribute	Apresenta	hasAttribute
Disease or Syndrome	DiagnosticadoCom	diagnosedWith
Invasive or Therapeutic Procedure	Recebe Medicacao Tratamento	receivesMedicationTreatment
Result	MonitoraPaciente	monitorsPatient
Sign or Symptom	ExibeSintoma	hasSymptom
Diagnostic Procedure	RealizaExame	performsTest
Clinical Note	TemNotaClinica	hasClinicalNote

Source: Created by the author.

Table 32: Object and Datatype Properties with their respective domains and ranges.

Property	**Domain**	**Range**
hasAttribute	Patient	ClinicalAttribute
hasSymptom	Patient	SignSymptom
diagnosedWith	Patient	DiseaseSyndrome
receivesMedicationTreatment	Patient	InvasiveTherapeuticProcedure
performsTest	Patient	DiagnosticProcedure
monitorsPatient	Patient	Result
hasClinicalNote	Patient	ClinicalNote
dateRecorded	ClinicalNote	xsd:dateTime

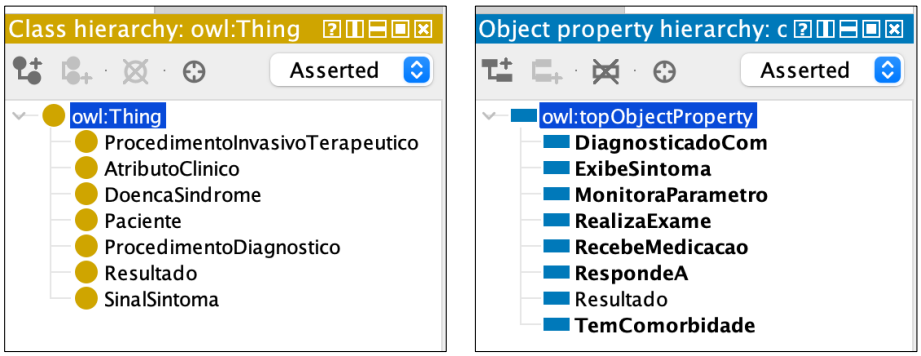
Source: Created by the author.

was establishing a relationship that specifies the domain and the range, providing explicit constraints and guidance on how data should be structured and interconnected within the ontology. Each class must be linked to the correct semantic information. This is guided by well-defined domain and range constraints, which play a pivotal role in maintaining the ontology's structural integrity and semantic clarity. These constraints are instrumental in preventing inappropriate connections between entities and ensuring that the data remains consistent and meaningful.

In OWL (Web Ontology Language), each property (relationship between classes) should specify a domain and a range. The domain indicates which class the property applies to, and the range defines the type of class that the property should point to. This allows the ontology to enforce data consistency rules by restricting how relationships can be used. We defined object properties, and their domain and range constraints are applied in our context, such as the conceptual definition in Table 32. Establishing the classes and relationships between them, as well as the constraints domain and range, we finalized the conceptual specification and requirements of our ontology structure.

Once our definitions and specifications for the ontology were finalized, we began with the classes and relationships that could adequately represent our data extracted from the

Figure 25: Initial draft of the proposed ontology with the relationship from classes to object properties



Source: Created by the author.

clinical notes, moving on to the ontology construction stage. During the second phase (2), we developed the ontology by creating the classes and relationships between the defined classes and structuring them into a coherent model. As observed in the clinical notes, these relationships capture the interactions between different clinical entities.

Considering the minimal specifications for our ontology, based on the essential data we extracted from the clinical notes, we developed the ontology using Protégé 5.6.4 (MUSEN, 2015), the latest version available. Protégé was employed to model and organize the concrete classes and relate the extracted data from the defined entities, ensuring that the semantic structure was suitable for consistently representing clinical information. In Figure 25, we show the core structure of our ontology, defining the classes and their relationships based on the extracted data from NER experiments to the ontology representation. A characteristic of the tool we used to define the concrete classes, such as Clinical Attributes, Invasive or Therapeutic Procedure, and Result, without spaces between entities or stop-words with no meaning like "OR," "AND" but keeping the aligning of ontology with the entities name.

These relationships were designed to reflect clinical processes and interactions, ensuring the ontology could accurately represent patient data, diagnosis, and treatment information, as presented in Figure 25. By formalizing these connections, we ensured that the ontology could support semantic reasoning and data integration, which are critical for the interoperability of health information systems. It reflects our efforts to structure ontology as patient-centric clinical data, ensuring each connection has a well-defined domain and range. This is crucial for maintaining data consistency and semantic clarity within the ontology. This design follows principles from leading interoperability standards, such as HL7 FHIR and OpenEHR, ensuring clinical information can be modeled, exchanged, and integrated across different systems.

We enable flexible reasoning and querying over the ontology by using object properties to link entities such as Patient, Disease or Syndrome, Result, Invasive or

Therapeutic Procedure, and ClinicalNote. For example, defining the '*performsTest*' property allows one to easily formulate queries about which tests were performed on a patient over a specific time period. Similarly, historical queries about treatments and outcomes can trace procedures through the '*receivesMedicationTreatment*' and '*monitorsPatient*' relationships.

In this sense, we can also highlight that our ontology's temporal characteristic allows us to reason about how a patient's condition has evolved (using properties like 'dataRecorded'). This is not just a theoretical concept but a practical tool that allows us to reason about the chronological progression of diseases, symptoms, and treatments by examining the timestamps associated with clinical notes and events. For example, we query to find all patients who were first diagnosed with COVID-19 (or any disease extracted from the clinical notes) and then later developed severe symptoms, such as dyspnea, requiring invasive treatment. This reasoning involves checking the order of events. A reasoner can infer if a patient's condition is worsening based on the sequence of '*diagnosedWith*', '*hasSymptom*', and '*receivesMedicationTreatment*' events over time.

Additionally, by assigning domains and ranges to each relationship, we ensure that the ontology maintains a clear structure, reducing potential ambiguities in how different data points relate. This structured approach supports and aligns with the broader goals of semantic interoperability, facilitating automated reasoning and enabling systems to infer additional relationships based on the data provided.

Following the third phase, we analyzed the structures of the published and selected ontologies in depth to understand which classes and/or relationships could be extended to them and reinforce our contextual meaning of extracted data. First, we began analyzing the COVID-19 Ontology, developed by Sargsyan et al. (2020), which presents a scope related to chemical entities cited for drug re-purposing and in the virus life cycle. It comprises 2270 classes of concepts and 38 987 axioms, tested on Medline. This approach enhances the ontology's ability to support semantic data extraction and retrieval, specifically targeting COVID-19-related information. From this ontology, we identified the class *COVID signs and symptoms* as a potential concrete alignment, offering a broad conceptual definition for our ontology:

> The *COVID signs and symptoms* concept enables a compilation of diverse concepts related to signs and symptoms of COVID-19 via the Axiome 'isSignsAndSymptomsFor some COVID-19'. This enables the adaptation of the ontology for text mining purposes for Covid-19-related information retrieval and information extraction. (SARGSYAN et al., 2020)

We also identified the possibility of extending the concept of our class ClinicalNote to the existing COVID clinical aspect class in the ontology, which, in its definition, includes a

description for representing clinical aspects related to our domain data. This class provides a structured way to capture and organize relevant clinical information, facilitating its use for tasks like semantic data extraction. By aligning our ClinicalNote class with this broader clinical aspect concept, we can enhance the representation of clinical observations and make our ontology more comprehensive for COVID-19 data analysis and retrieval:

> The *COVID clinical aspect* concept enables a compilation of diverse concepts related to clinical aspects in COVID-19-related research via the Axiome 'isClinicalAspectFor some COVID-19'. This enables the adaptation of the ontology for text mining purposes for Covid-19-related information retrieval and information extraction. (SARGSYAN et al., 2020)

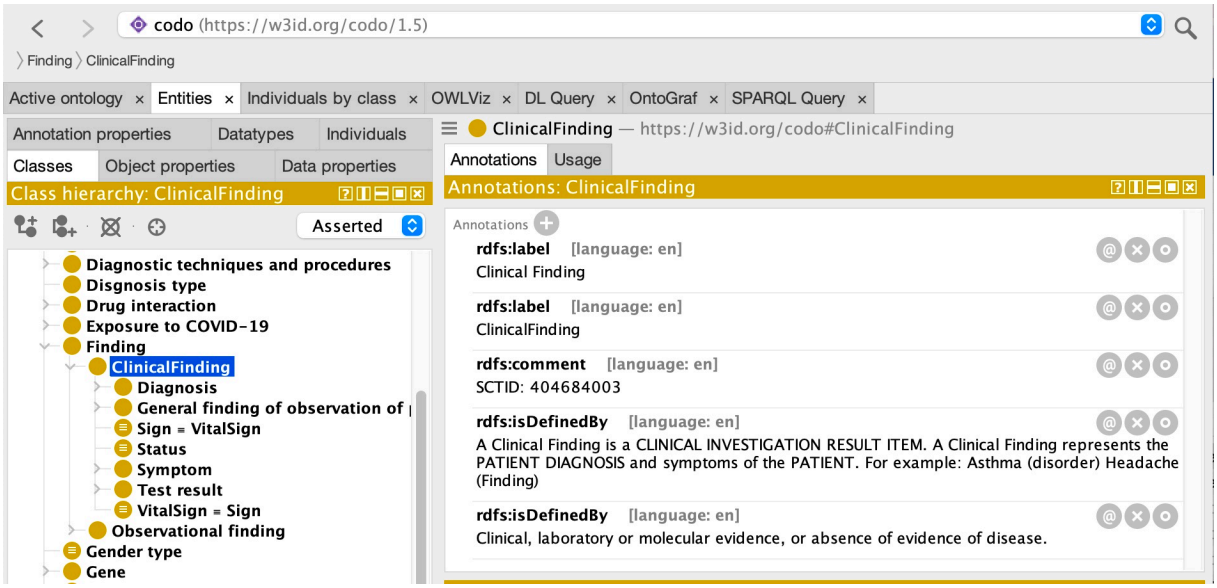
On the other hand, the ontology Coronavirus Infectious Disease Ontology (CIDO) developed by He et al. (2020), covers multiple domain of coronavirus diseases, including etiology, transmission, epidemiology, pathogenesis, diagnosis, prevention, and treatment. The authors later updated the ontology (HE et al., 2022) in a comprehensive study in 2022, focused on drugs, vaccines, and medical devices representations, composed of 6,758 terms, including imported terms from over 40 existing ontologies, and 190 CIDO-specific terms and hundreds of new axioms linking different entities. Some use cases CIDO was applied involved the Drug re-purposing by identifying 110 anti-coronavirus drugs. The ontology is available on BioPortal [8]

The COVID-19 Ontology for Data-driven Operationalization (CODO), developed by Dutta and DeBellis (2020), focuses on representing the operational aspects of COVID-19, including classes like patients, clinical findings, symptoms, and properties like relationships between patients, travel history, and test results. It comprises 385 classes and 16,682 axioms, enabling comprehensive semantic reasoning over COVID-19 data. One of the primary goals of CODO is to standardize how COVID-19 data is collected and represented across different datasets, ensuring that information about the pandemic can be easily shared and analyzed across systems. This ontology has been utilized in various data integration projects to support COVID-19 dashboards, which aggregate and present information on case counts, testing rates, and patient outcomes at the national and regional levels. CODO is available through the BioPortal repository[9], and Github repository repository[10]. The most relevant alignment we conducted with CODO ontology is related to three concrete Classes: Finding, Condition, Treatment procedure.

Similarly, the CovidO Ontology, created by Sharma and Jain (2024), targets the

[8]BioPortal CIDO<https://bioportal.bioontology.org/ontologies/CIDO/>
[9]BioPortal CODO<https://bioportal.bioontology.org/ontologies/CODO/>
[10]Github CODO<https://github.com/biswanathdutta/CODO>

Figure 26: Fragment of the concrete class Finding in CODO ontology, showing the possible alignments to our ontology.



Source: Created by the author based on CODO ontology structure.

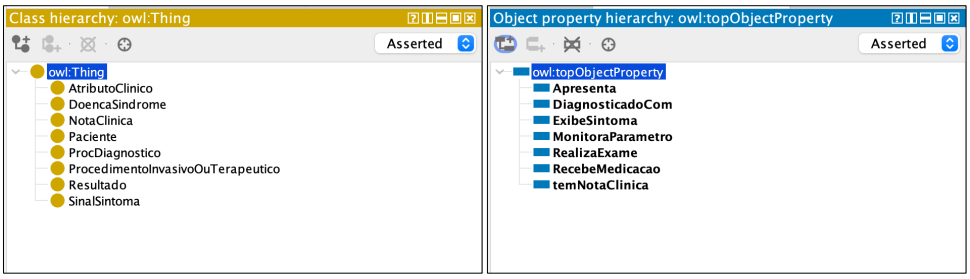
representation of clinical findings, laboratory results, and the progression of symptoms in COVID-19 patients. This ontology provides a structured framework to capture patient-specific data, which can then be used to analyze disease progression, treatment outcomes, and potential complications. CovidO consists of 823 classes and 12,506 axioms, with a strong focus on clinical outcomes and therapeutic interventions. CovidO also integrates terminology from existing clinical ontologies, including SNOMED CT and ICD-10, to ensure compatibility with broader medical datasets. This ontology has been applied in clinical decision support systems to aid healthcare providers in diagnosing and managing COVID-19 patients by offering a semantic framework for clinical decision-making. The ontology is accessible via BioPortal[11].

The fourth phase (4) involved populating the ontology with terms using SKOS (Simple Knowledge Organization System) to create thesauri aimed at semantic normalization on our extracted entities that presented a variability on terms. This process enabled us to standardize terminology across the dataset, ensuring that different terms referring to the same concept were unified under a single representation. By employing SKOS, we facilitated the normalization of clinical terms, improving the consistency and searchability of data across different systems and datasets. This semantic normalization plays a key role in enhancing interoperability, as it minimizes the variations in terminology that often arise in clinical texts.

Using an ontology structure to represent data is crucial in representing the key entities extracted from clinical notes in a semantic structure, an independent and

[11]CovidO<https://bioportal.bioontology.org/ontologies/COVIDO/>

Figure 27: Final ontology structure developed to represent the essential information from unstructured clinical notes.



Source: Created by the author.

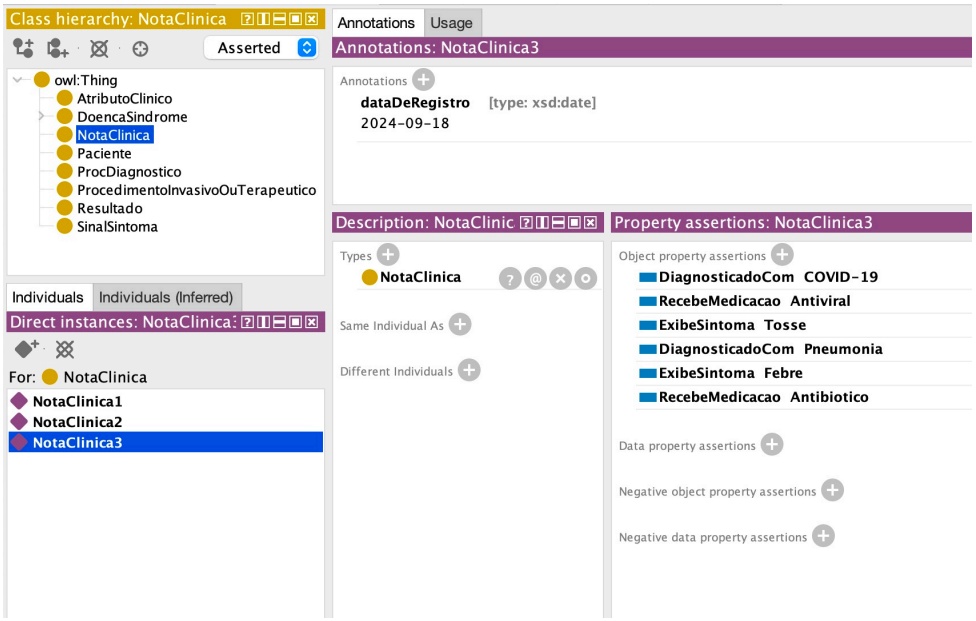
robust, ensuring a flexible and formal format. We carefully analyzed and harmonized our ontology with existent ontologies in the COVID-19 context, seeking to guarantee an alignment on validated and consolidated ontologies. This alignment ensures that our classes and properties are correlated with the main ontologies published and that the ontology is evaluated by ensuring the semantics and context already validated by specialists are semantically solid.

This assessment is vital as it verifies the integrity of the information represented correlated on ontologies validated by specialists, considering the high-level structured data requirements of a flexible structure such as ontology. To further ensure the feasibility and benefits of selecting an intermediary structure as an ontology, we evaluate its consistency by discussing the way and process to harmonize it with existing resources from the healthcare standard HL7 FHIR. This approach allowed us to verify that the data represented in the ontology meets the quality requirements of a healthcare standard, including compliance with terminologies, data types, and mandatory information from a formal protocol to standardize data.

Finally, we highlight from the developed ontology our systematic approach to define the structure based on the methodology that ensures each phase builds on the last, minimizing errors and aligning the ontology with real-world needs. We focused on interoperability aspects by aligning with existing COVID-19 ontologies. We sought to ensure that the ontology was internally consistent and capable of interoperating with other data models and systems. On the other hand, we also focused on integrating our NLP experiments and the ontology proposal by applying BERT models to extract data directly into the ontology, demonstrating a practical, scalable approach to populating and verifying the structure using unstructured clinical notes, as shown in Figure 28, where we bring a sample of three (3) clinical notes extracted and accommodated into the ontology.

In the final phase of the ontology development, we encountered lexical variability, a common issue when dealing with clinical data from unstructured sources like clinical notes. Lexical variability refers to using different terms, synonyms, or abbreviations

Figure 28: The developed ontology with a sample extract of clinical notes showing the semantic relation between data..



Source: Created by the author .

to describe the same concept. This can significantly hinder the healthcare domain's semantic consistency, data integration, and querying. Considering this challenge, we applied semantic normalization by leveraging the SKOS (Simple Knowledge Organization System) framework to address it. In the next section, we summarize the scenario and explain the approach to treating the situation.

6.5.1 Lexical normalization and semantic matching experiment

While we mentioned leveraging SKOS for managing lexical variations and synonyms, this process could become unmanageable in a large-scale clinical environment, where terminology is variable and evolving. For example, new terms related to diseases or treatments (especially in the context of rapidly evolving pandemics) will continuously emerge.

Our proposed solution to enhance the current problem considers employing ontological mappings to an external vocabulary manually built by the project's medicine students and specialists, such as for symptom, disease, and procedural terms. This allows us to inherit updates from these standard terminologies, and with further developing advances, we can add external mapping to international terminologies rather than manually handling synonym sets. By mapping to these large medical terminologies, we ensure that our ontology remains up-to-date and interoperable with a broad.

After constructing the ontology, one of the main challenges was ensuring the semantic consistency of terms extracted from clinical texts. Given the terminological

Figure 29: Semantic matching and lexical normalization mapping that the medical students developed to treat the variability

Source: Created by the author.

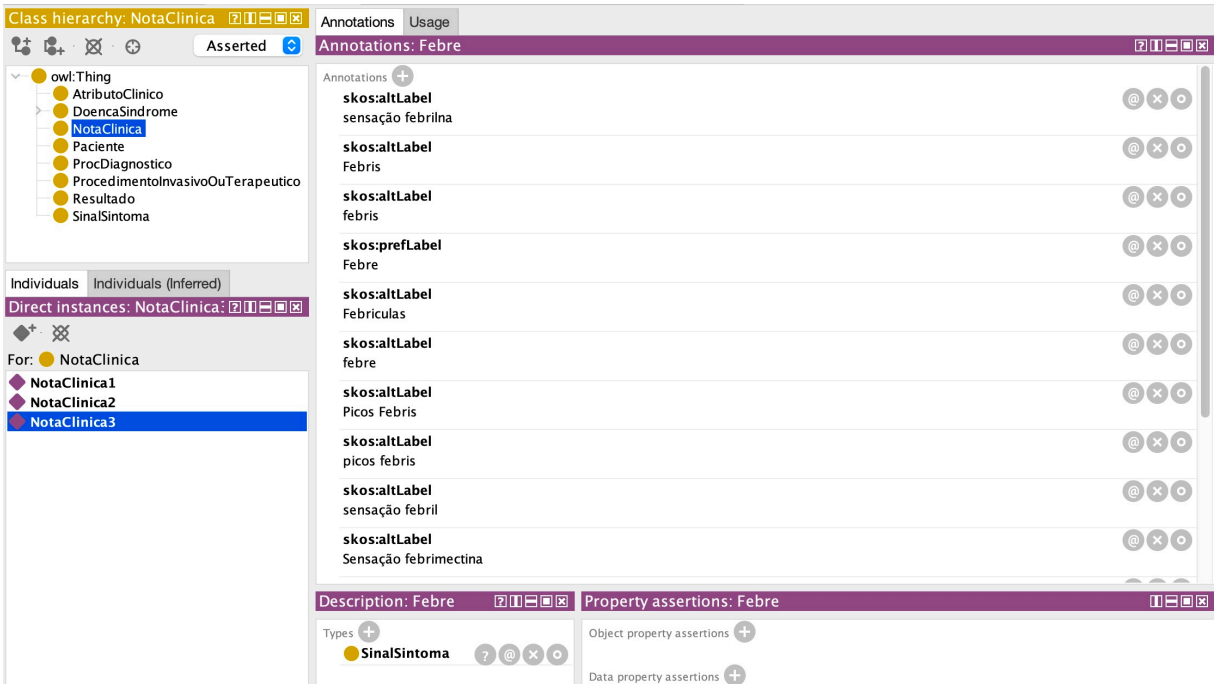
variation inherent in unstructured texts, such as clinical notes, we undertook a semantic normalization process using the SKOS (Simple Knowledge Organization System) standard. Applying SKOS helped organize the terms into a thesaurus specific to the clinical domain of COVID-19.

The mapped terms shown in Figure 29 were generated from a semantic analysis conducted by our medical students, whose practical knowledge of the COVID-19 context was crucial in identifying term variations and ensuring accuracy in the normalization process. By generating and mapping semantic entities based on our dataset, we delivered a dictionary containing the clinical entities identified through Named Entity Recognition (NER) experiments.

The semantic analysis of terms involved aligning clinical terms with their respective variations and synonyms, depending on the context in which the terms were used. For instance, terms like "COVID", "Covid", and "SARS-CoV" were grouped together, given their common reference to the viral infection.

We conducted a term grouping phase, focused on analyzing equivalent terms with different spellings or variants were grouped to ensure they were all normalized under the same class within the ontology. For example, terms like "Fever", "Febre" and "Febril" were mapped to the primary class of Fever. The creation of a COVID-19 thesaurus following the SKOS standard ensures that we are aligned with other ontologies standards to treat the thesaurus structure, it was structured to define term hierarchies and semantic relationships, improving interoperability and consistency in clinical

Figure 30: SKOS vocabulary incorporate on the proposed ontology as object annotations.



Source: Created by the author.

terminology throughout the data extraction process. In Figure 30 presents how we accommodate the mapped vocabulary at the ontology, observing the use of correct labels to identify a preferred concept and alternative concepts.

A critical part of creating our thesaurus involved aligning the terms with pre-existing clinical ontologies, such as CIDO, CODO, COVID-19 Ontology, and CovidO. This ensured that our vocabulary was interoperable with other international efforts. The alignment process included a detailed analysis of the term definitions in these recognized ontologies and matching them with the terms identified in our annotations. Ensuring compatibility with these ontologies guarantees that our ontology can easily integrate with other health data frameworks and medical ontologies.

One of the primary challenges was managing the insubstantial amount of terminological variation in unstructured clinical texts, especially for terms used ambiguously or with multiple meanings. The creation of a specific COVID-19 thesaurus was a valuable contribution to semantic normalization, enabling more accurate mapping and reducing ambiguities in clinical terminology. Additionally, this normalization greatly enhances the interoperability of clinical data across different healthcare systems, a critical factor in combating the pandemic.

Moreover, SKOS supports interoperability across different systems by adhering to a widely accepted standard for defining thesauri and concept labeling. This allows our ontology to integrate seamlessly with other systems or ontologies that also follow SKOS standards, empowering us to share and understand data across platforms. It facilitates

incorporation in a more extensive context, considering our objective of working in distributed clinical databases, such as the unique health system SUS. By incorporating SKOS, we achieve a more robust and consistent ontology and ensure that it can operate effectively within a broader ecosystem of healthcare systems. This enables enhanced data sharing and interoperability in alignment with global standards.

6.6 Discussion

During the COVID-19 pandemic, the rapid influx of new information and the urgency for quick decision-making posed significant challenges in processing clinical data. As noted by Raza and Schwartz (2023), the vast amounts of unstructured clinical data collected underscored the need for innovative approaches to managing this information. This unstructured data played a pivotal role in identifying new treatments and protocols. The demand for clinically relevant information highlighted the importance of efficient information extraction techniques, particularly in systematically recording positive cases and ensuring the clinical usefulness of collected data.

According to Paim (2018), the decentralization of resources and responsibilities in Brazil's Unified Health System (SUS) empowered the delivery of services and drove impactful health actions at all levels of care. This decentralization aligns with the constitutional guideline that establishes a unified command in each sphere of government. As a result, Brazil implemented a decentralized administration system across its 27 federal units and nearly 5,600 municipalities. Different government levels participate in defining public policies through conferences, councils, and commissions. This framework also supports the creation of agreements such as tripartite (federal, state, and municipal) and bipartite (state and municipal) inter-agency committees, which are responsible for coordinating the implementation of new public policies and guiding decisions in health governance. This decentralized structure allows for better resource allocation and more localized control over healthcare services, promoting tailored health interventions that respond to each region's specific needs.

In the Brazilian SUS management system, each healthcare unit and establishment can choose which electronic health systems to adopt for clinical information management. This autonomy contributes to the heterogeneity of the collected data, not only across different levels of care (primary care and specialized care) but also across various medical specialties. Furthermore, multiple proprietary systems, each with its database structure (LEE; KIM; LEE, 2021), exacerbate the challenges of uniformity and standardization in clinical data collection.

Recognizing the inevitability and necessity of adopting digital systems in the health domain, Brazil has proposed a national integrated network. In 2020, the Ministry of Health instituted the National Healthcare Data Network (RNDS), the national platform

to consolidate and reunite data from different healthcare public and private health establishments. Through Ordinance GM/MS Nº. 1,046 of May 24, 2021, during the pandemic, it was established the norm that mandates the reporting of COVID-19 Clinical Laboratory Diagnosis - Rapid Test Profiles results from public, private, university, and other laboratories across the national territory to the RNDS.

We explored the opportunity to work on an actual case study in partnership with the GHC hospital in Porto Alegre, with data from hospitalized patients who tested positive for COVID-19. Since the data represented a specialized reality, we developed a dataset as our first step in extracting and structuring data. We studied the data provided to understand which information in the clinical notes would be essential for extraction and structuring and which data to consider crucial so that the healthcare professional can monitor the patient's progress without spending time reading unstructured and complex textual data.

Observing the challenges of working with NER tasks, where we have a lack of annotated data in a universe of very specialized data, the authors Zha et al. (2024) suggest an approach that allows working with a hybrid approach, proposing the use of a small strongly labeled dataset combining the with a large weakly labeled data. Furthermore, they reinforce that using weakly labeled data does not necessarily improve performance; it can deteriorate the model's performance due to the extensive noise in the weak labels.

We looked at some studies that defined categories for unstructured health data to understand the granularity they applied and some characteristics in defining their particularities. Oliveira et al. (2020) performed the annotation in broader categories but with sub-annotations, which allows a greater granularity of the extracted information. Also reinforced by Zha et al. (2024), Jiang et al. (2021), NER tasks require specialized datasets, and usually, these tasks require token-level labels annotating a large number, which means a significant granularity.

The annotation team, composed of our medicine students and health specialists, supported the data annotation process. Clinical notes do not present a formal or solid structure, even though there is a standard for writing and collecting information in medical observations and clinical notes, called SOAP (subjective, objective, assessment, plan) (PODDER; LEW; GHASSEMZADEH, 2021), which we considered to clustering information when conducting the data analysis. The clinical notes analyzed follow in some way the method SOAP, providing a framework for evaluating information that healthcare professionals and system providers collect, a framework theorized and proposed by Larry Weed (WRIGHT et al., 2014).

Finally, we defined six (6) essential categories from the clinical notes around data we could extract: Clinical Attribute, Disease or Syndrome, Invasive Therapeutic Procedure, Result, Sign or Symptom, and Diagnostic Procedures. Understanding the information to

extract, we conducted the annotation process focusing on the categories defined and developed the COVID-19 dataset. Once the dataset was finalized, we executed our experiments to apply the NER task over the clinical notes data to evaluate and extract the essential information using BERT models (based on Transformer architecture). We have experienced three (3) models pre-trained on different data, a generalized multilingual model (Bert-base-multilingual-cased), a generalized Portuguese model (Bert-base-Portuguese-cased), and a domain-specific model (BioBERTpt-Clin) trained over electronic health records from a Brazilian hospital.

The categories we defined as essential information to extract from clinical notes were used to start the ontological structure definition since we aimed to represent the complete information extracted using the BERT model. The fully completed ontology structure comprised the categories and relationships between the extracted data. However, the decisions about the final representation in ontology were based on a previous study over four ontologies focused on COVID-19 data and the best practices of the healthcare standards openEHR and HL7 FHIR. It reflects our commitment to providing an ontology easily harmonized with healthcare standards to ensure structured data interoperability.

We represent the complete process of structuring unstructured data into ontologies by proposing a semantic interoperability model based on Natural Language Processing for non-structured health data. As a case study, the experiments presented in this research evaluate the proposed model for promoting the interoperability of unstructured clinical data. The limited dataset, constructed from COVID-19 clinical notes, demonstrated that it meets the basic needs for structuring and organizing clinical data, particularly managing the massive volume of unstructured data generated during the pandemic. Although the model was applied specifically to the COVID-19 context, its versatility and flexibility allow it to be expanded to other healthcare domains.

While the proposed model is not exclusively specialized in COVID-19, it is designed to be scalable across various healthcare areas, such as oncology, dentistry, radiology, endocrinology, and others. This scalability is made possible by the model's ability to integrate different domain-specific ontologies, enabling efficient organization and distribution of data representations according to the clinical context. Additionally, the proposed model is extensible, meaning it can easily be adapted to other types of unstructured data by creating new specialized datasets tailored to the specificity's of each medical field. The model's flexibility also lies in its distributed architecture, allowing different components to be developed and managed by multidisciplinary teams. This was demonstrated during the research development, in which experts from healthcare, software development, and artificial intelligence collaborated, proving that the model can be applied in various contexts and by professionals from different fields.

Practical applications include improved communication between healthcare providers

and the integration of electronic health records. The practical application of the proposed model has the potential to transform how healthcare professionals communicate and access relevant patient information, particularly in environments where clinical data is largely unstructured. By promoting semantic interoperability, the model automates the extraction of critical information from clinical records—such as diagnoses, treatments, and test results—regardless of the original format or text structure. This facilitates communication among healthcare professionals by ensuring that the relevant information is easily accessible and standardized.

As research contributions, we point out the proposed model for semantic interoperability for clinical notes, demonstrating high precision in entity extraction using BERT models in a specialized domain. The model ensured semantic consistency by aligning with COVID-19-specific ontologies like COVID-19 Ontology, COVID-19, CIDO, and CODO, addressing the lack of standardized healthcare terminologies in Portuguese and ontologies for clinical notes representation. Our research emphasizes this approach by directly focusing on real-world data, without undergoing prior treatment or enrichment during data extraction. This ensures a genuine application across different systems, considering the reality of unstructured data.

Regarding the flexible, scalable, and extensible characteristics, we highlight some model limitations, which include our case study, a regional focus on Porto Alegre, representing a small scope of Brazil data. Also, the absence of generalized terminologies that can be more or less applied in other clinical contexts requires the thesaurus to be updated and managed constantly. The dataset development and the need for specialists and experts to annotate and validate datasets when extending the model to other domains can overload and make projects more expensive. Although the unique structure of clinical notes in each hospital may challenge the model's broader application, we believe the necessary adjustments can be implemented with little effort. Handling that adaptability could make the model applicable to a broader range of hospitals and healthcare providers across Brazil. Adopting approaches that work directly with raw data without requiring significant modifications from healthcare providers and systems providers could significantly improve the quality of SUS data. Such strategies would enhance its utility for decision-making and secondary research and, most importantly, ensure continuity of patient care.

Future research directions involve expanding the model to other health domains, not only in medical clinical notes but also in nurse notes, financial notes, and administrative notes, ensuring the use of the collected information, even if it is unstructured. Focusing on integrating the model into healthcare systems with heterogeneous infrastructures, such as the Brazilian Unified Health System (SUS), which presents unique challenges due to the decentralized nature of its administration and diverse system providers. By working directly with raw data and aligning it with standardized ontologies and

healthcare standards like openEHR and HL7 FHIR, the model can enhance data quality, ensure interoperability, and ultimately improve the continuity of care. Another critical area for future research is the enhancement of semantic disambiguation. As noted, one of the challenges lies in aligning Portuguese healthcare terminologies with international standards like SNOMED-CT, LOINC, and ICD-10. These standards are widely adopted across international healthcare systems and offer a structured vocabulary for clinical terms, making global data exchange smoother. Finally, to achieve broader interoperability, we propose to equip the model with multi-lingual capabilities that handle the variability of medical terms in Portuguese and their alignment with international vocabularies. For instance, few-shot learning techniques could be employed to fine-tune models with smaller datasets in languages other than English or in specialized clinical domains, thus improving accuracy without requiring massive amounts of labeled data.

7 CONCLUSION

This research focused on to answer the follow research questions *How to **extract**, **disambiguate** and **represent** unstructured health data in an interoperability model using **hybrid approaches** combining natural language processing and machine learning techniques?*

Focusing on the research problem, in Chapter 3, we conducted a systematic review to explore gaps and opportunities to ensure interoperability on clinical data. The systematic review reinforced the challenges and difficulties of collecting and representing data from heterogeneous databases and standardizing data to allow interoperability. The growing demand for unstructured data in healthcare organizations has been a reality in recent years due to digital transformation, such as personal computers, devices, and the internet (HASAN; FARRI, 2019). In the healthcare domain, adopting EHR systems represents the transition from paper to digital (MARTIN-SANCHEZ; VERSPOOR, 2014). However, this rapid transition causes large volumes of data to be generated daily, with systems unprepared to receive complete information and palliative measures offering open textual fields for information insertion. It does so in an unstructured way, without controlled vocabularies, making sharing among institutions unfeasible.

In this scenario, and considering the gap and challenge involved in interoperating unstructured clinical data, in Chapter 4 we proposed a conceptual model to ensure semantic interoperability in unstructured clinical notes written in Portuguese, focusing on hybrids approaches, employing Natural Language Processing techniques with Machine Learning techniques to entity extraction. Our experiments demonstrated high precision, especially in extracting complex entities such as Diseases or Syndrome and Diagnostic Procedures, which are essential for analyzing medical records. This approach enhances the ability to process clinical data and addresses a critical gap: the absence of standardized and robust healthcare terminologies in Portuguese, a need identified in various studies (LANDOLSI; HLAOUA; ROMDHANE, 2023).

Our model stands out by improving the information extraction of essential terms from unstructured data, representing them into an semantic formal ontological structure, that can be shared among healthcare professionals and facilitating the integration with electronic health records, once it was harmonized with the main healthcare standards, HL7 FHIR and openEHR. Its practical application can be observed in the reduction of manual workload associated with reading and interpreting large volumes of textual data from clinical notes. It enables the automatic identification and extraction of relevant medical entities from clinical records, providing direct support for medical decision-making—an essential factor during health crises like the COVID-19 pandemic.

Our curated dataset was based on data from the Conceição Hospital Group (GHC), covering the period from December 28, 2020, to December 5, 2021. The dataset included

1,401 unique patients and 65,535 clinical note entries. Due to the unstructured nature of the data and the lack of consistent standards in Portuguese, we focused on annotating 314 documents, which resulted in the identification of 18,666 clinical entities. GHC, a reference in the Unified Health System (SUS), is the largest public hospital network in the south of Brazil and played a crucial role in providing healthcare services, especially during the pandemic.

Our work focused on hospitalized patients who tested positive for COVID-19, faithfully reflecting the reality of unstructured clinical data produced during one of the pandemic's most critical phases. As a result, the proposed model successfully captured the intricacies of medical records associated with COVID-19, aligning with specific ontologies like COVID-19, CIDO, and CODO to ensure semantic consistency. These efforts aimed to overcome the lack of standardized clinical note ontology that allows the representation of essential information for the Portuguese language, a recurring issue in applying NLP to healthcare data.

In conclusion, we proposed a semantic interoperability model to identify essential information within clinical notes for patients hospitalized with COVID-19. To achieve this, we developed an annotated dataset to facilitate the training of machine learning models, thereby automating the information extraction process. Once the dataset was curated and the models trained, we constructed an ontological structure to represent and organize the extracted data, providing a flexible intermediary structure. Understanding the lack of consensus on global interoperability standards, we focused on two of the most widely adopted standards: openEHR and HL7 FHIR. Our ontology was designed to align with both, ensuring that it can be easily adapted for future extensions. Additionally, we developed a lexical normalization and semantic alignment dictionary in Portuguese, which serves as a resource for enhancing the quality and consistency of the data. This dictionary can be published independently, allowing other projects to incorporate it into their ontologies or systems, promoting broader semantic interoperability in healthcare data management.

7.1 Contributions

This research had an exploratory and applied nature, observing the characteristics of real-world data generated daily by the routine of healthcare professionals in the context of the unified health system. The research contributes significantly to understanding how professionals' time can be optimized by allowing them to continue adopting the routine of clinical notes and collecting information in natural language. As well as ensuring that the rich data already collected can be extracted and organized in interoperable formats.

Also, in the context of the Brazilian healthcare system, which has the autonomy to adopt and contract different electronic medical record systems, our approach of structuring

the data in a formal ontology structure is highly practical. It allows each establishment, regardless of its technology choices, to easily adapt to the format, ensuring the feasibility of our proposed model in real-world healthcare settings.

The flexibility of an ontology for representing information ensures that the data can be harmonized to a health interoperability standard chosen by healthcare providers. As a contribution, we propose a model for implementing semantic interoperability, from the lowest and most technical level to the organizational level, that involves the governance of structured data.

As deliverables, we highlight the create dataset, that was developed by specialists and can be easily extended to improve performance to imbalance entities. We also provided a vocabulary of mapping COVID-19 lexical terms that can be extended to other ontologies and published separately. Ultimately, we reinforce the feasibility of the model to be generalized and implemented in other healthcare establishments of the unique health system context.

7.1.1 Published work

- **MELLO, B.H.**, RIGO, S.J., DA COSTA, C.A. et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol. (2022). https://doi.org/10.1007/s12553-022-00639-w

- **MELLO, B. H.**, DONIDA, B., DA COSTA, C. A., Rigo, S., ALVEZ, F., PETRY, L. R., ... DA ROSA RIGHI, R. (2021, June). SARS event notification: template openEHR brasileiro para interoperabilidade semântica de dados com foco na Covid-19. In Anais do XXI Simpósio Brasileiro de Computação Aplicada à Saúde (pp. 434-439). SBC. https://doi.org/10.5753/sbcas.2021.16088.

- **MELLO, B.H.** et al. SARS event notification. 13 Oct. 2020. Available from: https://ckm.openehr.org/ckm/templates/1013.26.377. Accessed: 1 Mar. 2022.

In addition, several publications were developed in the context of the project and in collaboration with other researchers. The following list presents the other publications:

- Published articles:

  - SCHÜNKE, L. C.; **MELLO, B. H.**; COSTA, C. A.; ANTUNES, R. S.; RIGO, S. J.; RAMOS, G. O.; RIGHI, R. R.; SCHERER, J.N.; DONIDA, B. A Rapid Review of Machine Learning Approaches for Telemedicine in the Scope of COVID-19. Artificial Intelligence in Medicine, 2022. https://doi.org/10.1016/j.artmed.2022.102312

  - SILVA, D.P.; **MELLO, B.H.**; FRÖHLICH, W.R.; RIGO, S.J.; SCHWERTNER, M.A.; COSTA, C.A. Avaliação de modelos para extração de dados não estruturados de um sistema EHR para atender a estrutura final de uma ontologia. In: SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO APLICADA À SAÚDE (SBCAS), 22. , 2022, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2022.
  - DONIDA, B.; **MELLO, B. H.**; ALVES, F. H. ; RODRIGUES, H. J. G. C. ; REISDORFER, L. F. ; MULLER, A. P. ; ALEGRETTI, A. P. ; SCHERER, J. N. ; RIGO, S. J. ; COSTA, C. A. . Primeiro Arquétipo Brasileiro na Plataforma OpenEhr para Interoperabilidade Semântica de Dados com foco na Covid-19. In: 40ª Semana Científica do Hospital de Clínicas de Porto Alegre, 2020, Porto Alegre. Anais da 40ª Semana Científica do Hospital de Clínicas de Porto Alegre., 2020. v. 1. p. 1-2.
  - CABRAL, A., FISCHER, G., RIGHI, R., COSTA, C., ROEHRS, A., RIGO, S., and MELLO, B. (2024). Providing Interoperability between Wearable Devices and FHIR-based Healthcare Systems. In Proceedings of the 30th Brazilian Symposium on Multimedia and the Web, (pp. 410-414). Porto Alegre: SBC. doi:10.5753/webmedia.2024.243230

- Published book chapters:

  - COSTA, C. A. ; ZEISER, F. A. ; RIGHI, R. R. ; ANTUNES, R. S. ; ALEGRETTI, A. P. ; BERTONI, A. P. ; RAMOS, G. O. ; **MELLO, B. H.** ; VANIN, F. ; BERTOLETTI, O. A. ; RIGO, S. J. . Internet of Things and Machine Learning for Smart Healthcare. 2024.
  - BERTONI, A. P. S. ; RODRIGUES, V. F. ; ZEISER, F. A. ; **MELLO, B. H.** ; COSTA, C. A. ; DONIDA, B. ; RIGO, S. J. ; RIGHI, R. R. . Internet das Coisas de Saúde: aplicando IoT, interoperabilidade e aprendizado de máquina com foco no paciente. Minicursos do XXII Simpósio Brasileiro de Computação Aplicada à Saúde. 1ed.: , 2022, v. , p. 1-. DOI: 10.5753/sbc.10508.8

7.2 Limitations

The proposed model for achieving semantic interoperability in unstructured clinical data has a conceptual foundation, and we developed a prototype to evaluate the feasibility of its main components in real-world scenarios. One of the most significant limitations is the data's heterogeneity and inconsistency, which poses challenges for training machine learning (ML) models. Many preprocessing steps are required to make the data suitable for model training, which adds complexity to the implementation.

An essential limitation of the study is its regional focus because the case study was applied to data exclusively collected from the Conceição Hospital Group (GHC) in Porto Alegre, Rio Grande do Sul (RS). While GHC is the largest public hospital network in southern Brazil and serves a broad patient base, the lack of detailed demographic information about patients' geographical origins limits the generalizability of the findings to other regions in Brazil. Additionally, the clinical note structures reflect a regional healthcare scenario influenced by the specific documentation practices and terminologies used in RS's public health system. These practices may not directly translate to other regional or international healthcare contexts. Also, the annotation and validation of unstructured clinical data necessitate the involvement of medical experts, which may limit the model's scalability to hospital environments that use different documentation practices and terminologies.

Future implementations will likely need to adapt to the heterogeneity of healthcare systems across Brazil and beyond to apply the model more broadly. This underscores the importance of aligning the model with more universal healthcare standards that accommodate different regional practices.

The dataset publication restrictions represent a limitation due to ethics and institutional research committee guidelines. The dataset used in this study cannot be publicly shared without formal authorization from the hospital partner. However, we consider the data annotation methodology and the use of advanced tools a valuable contribution to the field.

Regarding lexical normalization, while manual lexical normalization performed by specialists improved the results, it required considerable effort, which poses scalability challenges. Automated methods may be needed to scale the process across larger datasets.

We applied a domain-specific dataset. The annotated dataset is specific to COVID-19 inpatient data, limiting the model's generalization to broader healthcare domains. Future studies should aim to expand this model to more generalized health data.

7.3 Future work

Expanding the developed ontological structure: A key direction for future research involves expanding the ontological structure developed in this study to encompass additional aspects of patient clinical data. This could include demographic information, immunization records, laboratory results, and drug interactions. By extending the ontology in these areas, the model could support a more comprehensive application across various critical healthcare domains. This broader scope would facilitate more effective data interoperability between clinical systems and specialties, enhancing patient care and medical research. Incorporating these additional data points would also allow for more detailed patient histories and improved clinical decision-making.

Lexical Normalization and Alignment with International Terminologies: A structured and consistent framework for lexical normalization is essential to enhance the model's applicability further. Aligning this framework with international medical terminologies like SNOMED-CT, LOINC, and ICD-10 would ensure the model can standardize medical terms across different languages and contexts. This alignment would promote global interoperability, enabling the seamless sharing of clinical information across borders and improving the accuracy of information extraction and usage in various healthcare systems.

Semantic Disambiguation with Graph Neural Networks (GNNs): Another future exploration is utilizing Graph Neural Networks (GNNs) to improve semantic disambiguation and term correlation within the thesaurus developed in this study. The ontology created here could serve as a training dataset for GNNs, which can be applied to align extracted clinical terms with existing knowledge bases. Similar approaches have successfully aligned international terminologies and integrated diverse knowledge sources (ZHOU et al., 2020).

Integration with New Clinical Domains: Finally, the model's scalability could be tested by expanding its application to other health domains beyond COVID-19, such as oncology, dentistry, endocrinology, and radiology. This expansion would test the model's scalability and adaptability in new clinical contexts and contribute to developing specialized datasets tailored to these areas. By adapting the model to handle unstructured data from various medical fields, future research can further validate its generalizability and ensure that it can be applied across various healthcare systems and specialties. This expanded scope, combined with enhancements in semantic disambiguation and international alignment, would make the model a valuable tool for advancing semantic interoperability and improving the quality of clinical data management.

REFERENCES

ABDELGHANY, Abdelghany Salah; DARWISH, Nagy Ramadan; HEFNI, Hesham Ahmed. An agile methodology for ontology development. **International Journal of Intelligent Engineering and Systems**, v. 12, n. 2, p. 170–181, 2019.

ADEL, Ebtsam et al. Ontology-based electronic health record semantic interoperability: A survey. In: **U-Healthcare Monitoring Systems**. [S.l.]: Elsevier, 2019. p. 315–352.

AGRAWAL, Rashmi et al. Machine learning for healthcare–1.fundamentals of machine learning. Chapman and Hall/CRCR, v. 1st Edition, n. 1, p. 5–18, 2020.

ALEXOPOULOS, Panos. **Semantic Modeling for Data**. 1st ed.. ed. [S.l.]: O'Reilly Media, 2020.

ALLEN, James F. Natural language processing. In: **Encyclopedia of computer science**. [S.l.: s.n.], 2003. p. 1218–1222.

AMMENWERTH, Elske et al. The effect of electronic prescribing on medication errors and adverse drug events: a systematic review. **Journal of the American Medical Informatics Association**, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 15, n. 5, p. 585–600, 2008.

ASSOCIATION, National Electrical Manufacturers et al. Digital imaging and communications in medicine (dicom). **https://www.dicomstandard.org/**, 2003.

BAE, Sungchul; YI, Byoung-Kee. Development of eclaim system for private indemnity health insurance in south korea: Compatibility and interoperability. **Health Informatics Journal**, SAGE Publications Sage UK: London, England, v. 28, n. 1, p. 14604582211071019, 2022.

BAHGA, Arshdeep; MADISETTI, Vijay K. A cloud-based approach for interoperable electronic health records (ehrs). **IEEE Journal of Biomedical and Health Informatics**, IEEE, v. 17, n. 5, p. 894–906, 2013.

BEALE, Thomas; HEARD, Sam et al. An ontology-based model of clinical information. **Studies in health technology and informatics**, IOS Press; 1999, v. 129, n. 1, p. 760, 2007.

BEAULIEU-JONES, Brett K; GREENE, Casey S et al. Semi-supervised learning of the electronic health record for phenotype stratification. **Journal of biomedical informatics**, Elsevier, v. 64, p. 168–178, 2016.

BENGIO, Yoshua; LECUN, Yann; HINTON, Geoffrey. Deep learning for ai. **Communications of the ACM**, ACM New York, NY, USA, v. 64, n. 7, p. 58–65, 2021.

BENSON, Tim; GRIEVE, Grahame. Why interoperability is hard. In: **Principles of Health Interoperability**. [S.l.]: Springer, 2021. p. 21–40.

BLACKMAN-LEES, Shellon. Towards a conceptual framework for persistent use: A technical plan to achieve semantic interoperability within electronic health record systems. In: **Proceedings of the 51st Hawaii International Conference on System Sciences**. [S.l.: s.n.], 2018.

BLOOMROSEN, Meryl; BERNER, Eta S et al. Findings from the health information management section of the 2020 international medical informatics association yearbook. **Yearbook of Medical Informatics**, Georg Thieme Verlag KG, v. 29, n. 01, p. 087–092, 2020.

BONO, Bernard De et al. The ricordo approach to semantic interoperability for biomedical data and models: strategy, standards and solutions. **BMC Research Notes**, BioMed Central, v. 4, n. 1, p. 1–15, 2011.

BOUARROUDJ, Wissem; BOUFAIDA, Zizette; BELLATRECHE, Ladjel. Named entity disambiguation in short texts over knowledge graphs. **Knowledge and Information Systems**, Springer, p. 1–27, 2022.

COMMISSION, Europian et al. European interoperability framework–implementation strategy. **COM (2017) 134 final**, 2017.

DASH, Sabyasachi et al. Big data in healthcare: management, analysis and future prospects. **Journal of Big Data**, SpringerOpen, v. 6, n. 1, p. 1–25, 2019.

DEMSKI, Hans; GARDE, Sebastian; HILDEBRAND, Claudia. Open data models for smart health interconnected applications: the example of openehr. **BMC medical informatics and decision making**, Springer, v. 16, n. 1, p. 1–9, 2016.

DEVLIN, Jacob et al. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

DUFTSCHMID, Georg; CHALOUPKA, Judith; RINNER, Christoph. Towards plug-and-play integration of archetypes into legacy electronic health record systems: the archimed experience. **BMC medical informatics and decision making**, BioMed Central, v. 13, n. 1, p. 1–12, 2013.

DUFTSCHMID, Georg; WRBA, Thomas; RINNER, Christoph. Extraction of standardized archetyped data from electronic health record systems based on the entity-attribute-value model. **International journal of medical informatics**, Elsevier, v. 79, n. 8, p. 585–597, 2010.

DUTTA, Biswanath; DEBELLIS, Michael. Codo: an ontology for collection and analysis of covid-19 data. **arXiv preprint arXiv:2009.01210**, 2020.

D'AMORE, John D et al. Clinical data sharing improves quality measurement and patient safety. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 28, n. 7, p. 1534–1542, 2021.

ELLOUZE, Afef Samet; BOUAZIZ, Rafik; GHORBEL, Hanen. Integrating semantic dimension into openehr archetypes for the management of cerebral palsy electronic medical records. **Journal of biomedical informatics**, Elsevier, v. 63, p. 307–324, 2016.

EVANS, R Scott. Electronic health records: then, now, and in the future. **Yearbook of medical informatics**, Georg Thieme Verlag KG, v. 25, n. S 01, p. S48–S61, 2016.

FERNÁNDEZ-BREIS, Jesualdo Tomás et al. Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. **Journal of the American Medical Informatics Association**, BMJ Publishing Group BMA House, Tavistock Square, London, WC1H 9JR, v. 20, n. e2, p. e288–e296, 2013.

FOREMAN, Caroline et al. Categorization of adverse drug reactions in electronic health records. **Pharmacology Research & Perspectives**, Wiley Online Library, v. 8, n. 2, p. e00550, 2020.

FU, Sunyang et al. Natural language processing for the evaluation of methodological standards and best practices of ehr-based clinical research. **AMIA Summits on Translational Science Proceedings**, American Medical Informatics Association, v. 2020, p. 171, 2020.

GAGALOVA, Kristina K et al. What you need to know before implementing a clinical research data warehouse: comparative review of integrated data repositories in health care institutions. **JMIR formative research**, JMIR Publications Inc., Toronto, Canada, v. 4, n. 8, p. e17687, 2020.

GANSEL, Xavier; MARY, Melissa; BELKUM, Alex van. Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review. **European Journal of Clinical Microbiology & Infectious Diseases**, Springer, v. 38, n. 6, p. 1023–1034, 2019.

GIBAUD, Bernard. The dicom standard: a brief overview. **Molecular imaging: computer reconstruction and practice**, Springer, p. 229–238, 2008.

GIORGI, John M; BADER, Gary D. Towards reliable named entity recognition in the biomedical domain. **Bioinformatics**, Oxford University Press, v. 36, n. 1, p. 280–286, 2020.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep learning**. [S.l.]: MIT press, 2016.

GRUBER, Thomas R. A translation approach to portable ontology specifications. **Knowledge acquisition**, Elsevier, v. 5, n. 2, p. 199–220, 1993.

GRUBER, Thomas R. Toward principles for the design of ontologies used for knowledge sharing? **International Journal of Human-Computer Studies**, v. 43, n. 5, p. 907–928, 1995. ISSN 1071-5819.

GUPTA, Pankaj et al. Neural relation extraction within and across sentence boundaries. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2019. v. 33, n. 01, p. 6513–6520.

HAKALA, Kai; PYYSALO, Sampo. Biomedical named entity recognition with multilingual bert. p. 56–61, 2019.

HASAN, Sadid A; FARRI, Oladimeji. Clinical natural language processing with deep learning. In: **Data science for healthcare**. [S.l.]: Springer, 2019. p. 147–171.

HASSANPOUR, Saeed; LANGLOTZ, Curtis P. Information extraction from multi-institutional radiology reports. **Artificial intelligence in medicine**, Elsevier, v. 66, p. 29–39, 2016.

HAYMAN, James A et al. Minimum data elements for radiation oncology: An american society for radiation oncology consensus paper. **Practical radiation oncology**, Elsevier, v. 9, n. 6, p. 395–401, 2019.

HE, Yongqun et al. A comprehensive update on cido: the community-based coronavirus infectious disease ontology. **Journal of biomedical semantics**, Springer, v. 13, n. 1, p. 25, 2022.

_____. Cido, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. **Scientific data**, Nature Publishing Group UK London, v. 7, n. 1, p. 181, 2020.

HERRMANN, Markus D et al. Implementing the dicom standard for digital pathology. **Journal of pathology informatics**, Wolters Kluwer–Medknow Publications, v. 9, 2018.

HEYMANN, David L; SHINDO, Nahoko. Covid-19: what is next for public health? **The lancet**, Elsevier, v. 395, n. 10224, p. 542–545, 2020.

HIMSS. **Healthcare Information and Management Systems Society**. [S.l.], 2021. Available at: <https://www.himss.org/resources/interoperability-healthcare>.

HITZLER, Pascal. A review of the semantic web field. **Communications of the ACM**, ACM New York, NY, USA, v. 64, n. 2, p. 76–83, 2021.

HONG, Lixiang et al. A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories. **Nature Machine Intelligence**, Nature Publishing Group, v. 2, n. 6, p. 347–355, 2020.

HOUSSEIN, Essam H; MOHAMED, Rehab E; ALI, Abdelmgeid A. Machine learning techniques for biomedical natural language processing: A comprehensive review. **IEEE Access**, IEEE, 2021.

HULSEN, Tim. Sharing is caring—data sharing initiatives in healthcare. **International journal of environmental research and public health**, MDPI, v. 17, n. 9, p. 3046, 2020.

ISO13606. **ISO 13606:2019 Standard - EHR Interoperability**. [S.l.], 2019.

ISO18308. **ISO 18308:2011(en), Health informatics — Requirements for an Electronic Health Record Architecture**. 2011. Available at: <https://www.iso.org/obp/ui/#iso:std:iso:18308:ed-1:v1:en>.

ISOTR20514. **ISO 20514:2005 Health informatics - electronic health record - definition, scope and context**. [S.l.], 2005. 6-8 p.

JI, Zongcheng; WEI, Qiang; XU, Hua. Bert-based ranking for biomedical entity normalization. **AMIA Summits on Translational Science Proceedings**, American Medical Informatics Association, v. 2020, p. 269, 2020.

JIANG, Haoming et al. Named entity recognition with small strongly labeled and large weakly labeled data. **arXiv preprint arXiv:2106.08977**, 2021.

JORDAN, Michael I; MITCHELL, Tom M. Machine learning: Trends, perspectives, and prospects. **Science**, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015.

JR, W Dean Bidgood et al. Understanding and using dicom, the data interchange standard for biomedical imaging. **Journal of the American Medical Informatics Association**, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 4, n. 3, p. 199–212, 1997.

KAH, Anoual El; ZEROUAL, Imad. A review on applied natural language processing to electronic health records. In: IEEE. **2021 1st International Conference on Emerging Smart Technologies and Applications (eSmarTA)**. [S.l.], 2021. p. 1–6.

KALOGIANNIS, Spyridon et al. Integrating an openehr-based personalized virtual model for the ageing population within hbase. **BMC medical informatics and decision making**, BioMed Central, v. 19, n. 1, p. 1–15, 2019.

KASTHURIRATHNE, Suranga N et al. Enabling better interoperability for healthcare: lessons in developing a standards based application programing interface for electronic medical record systems. **Journal of medical systems**, Springer, v. 39, n. 11, p. 1–8, 2015.

KIM, Dae-young; JOSHI, Karuna P. A semantically rich knowledge graph to automate hipaa regulations for cloud health it services. In: IEEE. **2021 7th IEEE Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)**. [S.l.], 2021. p. 7–12.

KITCHENHAM, Barbara; BRERETON, Pearl. A systematic review of systematic review process research in software engineering. **Information and software technology**, Elsevier, v. 55, n. 12, p. 2049–2075, 2013.

KOLECK, Theresa A et al. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 26, n. 4, p. 364–379, 2019.

KOLUKULA, Nitalaksheswara Rao et al. Processing of clinical notes for efficient diagnosis with feedback attention–based bilstm. **Medical & Biological Engineering & Computing**, Springer, p. 1–16, 2024.

KONSTANTINOVA, Natalia. Review of relation extraction methods: What is new out there? p. 15–28, 2014.

KOPANITSA, Georgy. Integration of hospital information and clinical decision support systems to enable the reuse of electronic health record data. **Methods of information in medicine**, Schattauer GmbH, v. 56, n. 03, p. 238–247, 2017.

KORMILITZIN, Andrey et al. Med7: a transferable clinical natural language processing model for electronic health records. **Artificial Intelligence in Medicine**, Elsevier, v. 118, p. 102086, 2021.

KUSHIDA, Clete A et al. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. **Medical care**, NIH Public Access, v. 50, n. Suppl, p. S82, 2012.

LAAR, Sylvia A van et al. An electronic health record text mining tool to collect real-world drug treatment outcomes: A validation study in patients with metastatic renal cell carcinoma. **Clinical Pharmacology & Therapeutics**, Wiley Online Library, v. 108, n. 3, p. 644–652, 2020.

LAMPRINAKOS, Georgios C et al. Using fhir to develop a healthcare mobile application. In: IEEE. **2014 4th international conference on wireless mobile communication and healthcare-transforming healthcare through innovations in mobile and wireless technologies (MOBIHEALTH)**. [S.l.], 2014. p. 132–135.

LANDOLSI, Mohamed Yassine; HLAOUA, Lobna; ROMDHANE, Lotfi Ben. Information extraction from electronic medical documents: state of the art and future research directions. **Knowledge and Information Systems**, Springer, v. 65, n. 2, p. 463–516, 2023.

LAPARRA, Egoitz et al. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. **Yearbook of Medical Informatics**, Georg Thieme Verlag KG, v. 30, n. 01, p. 239–244, 2021.

LAZAROVA, Elena et al. An interoperable electronic health record system for clinical cardiology. In: MDPI. **Informatics**. [S.l.], 2022. v. 9, n. 2, p. 47.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

LEE, Ah Ra; KIM, Il Kon; LEE, Eunjoo. Developing a transnational health record framework with level-specific interoperability guidelines based on a related literature review. In: MULTIDISCIPLINARY DIGITAL PUBLISHING INSTITUTE. **Healthcare**. [S.l.], 2021. v. 9, n. 1, p. 67.

LEGAZ-GARCÍA, María del Carmen et al. Transformation of standardized clinical models based on owl technologies: from cem to openehr archetypes. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 22, n. 3, p. 536–544, 2015.

LEGAZ-GARCÍA, María del Carmen et al. A semantic web based framework for the interoperability and exploitation of clinical models and ehr data. **Knowledge-Based Systems**, Elsevier, v. 105, p. 175–189, 2016.

LEZCANO, Leonardo; SICILIA, Miguel-Angel; RODRÍGUEZ-SOLANO, Carlos. Integrating reasoning and clinical archetypes using owl ontologies and swrl rules. **Journal of biomedical informatics**, Elsevier, v. 44, n. 2, p. 343–353, 2011.

LI, Fei et al. Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: an empirical study. **JMIR medical informatics**, JMIR Publications Inc., Toronto, Canada, v. 7, n. 3, p. e14830, 2019.

LI, Irene et al. Neural natural language processing for unstructured data in electronic health records: a review. **arXiv preprint arXiv:2107.02975**, 2021.

LI, Jing et al. A survey on deep learning for named entity recognition. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, 2020.

LI, Mengyang et al. Development of an openehr template for covid-19 based on clinical guidelines. **Journal of medical Internet research**, JMIR Publications Inc., Toronto, Canada, v. 22, n. 6, p. e20239, 2020.

LOPES, Fábio; TEIXEIRA, César; OLIVEIRA, Hugo Gonçalo. Contributions to clinical named entity recognition in portuguese. In: **Proceedings of the 18th BioNLP Workshop and Shared Task**. [S.l.: s.n.], 2019. p. 223–233.

LOZANO-RUBÍ, Raimundo et al. Ontocr: A cen/iso-13606 clinical repository based on ontologies. **Journal of biomedical informatics**, Elsevier, v. 60, p. 224–233, 2016.

LUOMA, Jouni; PYYSALO, Sampo. Exploring cross-sentence contexts for named entity recognition with bert. **arXiv preprint arXiv:2006.01563**, 2020.

MA, Kai et al. Ontology-based bert model for automated information extraction from geological hazard reports. **Journal of Earth Science**, Springer, v. 34, n. 5, p. 1390–1405, 2023.

MALDONADO, José Alberto et al. Clin-ik-links: a platform for the design and execution of clinical data transformation and reasoning workflows. **Computer Methods and Programs in Biomedicine**, Elsevier, v. 197, p. 105616, 2020.

MALM-NICOLAISEN, Kristian et al. Efforts on using standards for defining the structuring of electronic health record data: A scoping review. In: LINKÖPING UNIVERSITY ELECTRONIC PRESS. **SHI 2019. Proceedings of the 17th Scandinavian Conference on Health Informatics, November 12-13, 2019, Oslo, Norway**. [S.l.], 2019. p. 108–115.

MANNING, Christopher; SCHUTZE, Hinrich. **Foundations of statistical natural language processing**. [S.l.]: MIT press, 1999.

MARCO-RUIZ, Luis et al. Archetype-based data warehouse environment to enable the reuse of electronic health record data. **International journal of medical informatics**, Elsevier, v. 84, n. 9, p. 702–714, 2015.

MARCOS, Carlos et al. Solving the interoperability challenge of a distributed complex patient guidance system: a data integrator based on hl7's virtual medical record standard. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 22, n. 3, p. 587–599, 2015.

MARGHERI, Andrea et al. Decentralised provenance for healthcare data. **International Journal of Medical Informatics**, Elsevier, v. 141, p. 104197, 2020.

MARTIN-SANCHEZ, Fernando; VERSPOOR, Karin. Big data in medicine is driving big changes. **Yearbook of medical informatics**, Georg Thieme Verlag KG, v. 23, n. 01, p. 14–20, 2014.

MARTÍNEZ-COSTA, Catalina et al. Semantic enrichment of clinical models towards semantic interoperability. the heart failure summary use case. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 22, n. 3, p. 565–576, 2015.

MARTÍNEZ-COSTA, Catalina; MENÁRGUEZ-TORTOSA, Marcos; FERNÁNDEZ-BREIS, Jesualdo Tomás. An approach for the semantic interoperability of iso en 13606 and openehr archetypes. **Journal of biomedical informatics**, Elsevier, v. 43, n. 5, p. 736–746, 2010.

MENÁRGUEZ-TORTOSA, Marcos; MARTÍNEZ-COSTA, Catalina; FERNÁNDEZ-BREIS, Jesualdo Tomás. A generative tool for building health applications driven by iso 13606 archetypes. **Journal of medical systems**, Springer, v. 36, n. 5, p. 3063–3075, 2012.

MENDES, David et al. Enrichment/population of customized cpr (computer-based patient record) ontology from free-text reports for csi (computer semantic interoperability). **Procedia Technology**, Elsevier, v. 5, p. 753–762, 2012.

MILDENBERGER, Peter; EICHELBERG, Marco; MARTIN, Eric. Introduction to the dicom standard. **European radiology**, Springer, v. 12, n. 4, p. 920–927, 2002.

MIN, Lingtong et al. An openehr based approach to improve the semantic interoperability of clinical data registry. **BMC medical informatics and decision making**, BioMed Central, v. 18, n. 1, p. 49–56, 2018.

MITCHELL, Tom M. Does machine learning really work? **AI magazine**, v. 18, n. 3, p. 11–11, 1997.

MITCHELL, Tom M. **Machine Learning**. Fourth. Upper Saddle River: McGraw-Hill Science, 1997. 892 p.

MOREIRA, Mario WL et al. Semantic interoperability and pattern classification for a service-oriented architecture in pregnancy care. **Future Generation Computer Systems**, Elsevier, v. 89, p. 137–147, 2018.

MORO, Andrea; RAGANATO, Alessandro; NAVIGLI, Roberto. Entity linking meets word sense disambiguation: a unified approach. **Transactions of the Association for Computational Linguistics**, MIT Press, v. 2, p. 231–244, 2014.

MOUGIN, Fleur; HOLLIS, Kate Fultz; SOUALMIA, Lina F. Inclusive digital health. **Yearbook of Medical Informatics**, Georg Thieme Verlag KG, v. 31, n. 01, p. 002–006, 2022.

MUSEN, Mark A. The protégé project: a look back and a look forward. **AI matters**, ACM New York, NY, USA, v. 1, n. 4, p. 4–12, 2015.

NAQA, Issam El; MURPHY, Martin J. What is machine learning? In: **machine learning in radiation oncology**. [S.l.]: Springer, 2015. p. 3–11.

NASAR, Zara; JAFFRY, Syed Waqar; MALIK, Muhammad Kamran. Named entity recognition and relation extraction: State-of-the-art. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 54, n. 1, p. 1–39, 2021.

NASEEM, Usman et al. Biomedical named-entity recognition by hierarchically fusing biobert representations and deep contextual-level word-embedding. p. 1–8, 2020.

'NASEEM, Usman' et al. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. **Transactions on Asian and Low-Resource Language Information Processing**, ACM New York, v. 20, n. 5, p. 1–35, 2021.

NEGRO-CALDUCH, Elsa et al. Technological progress in electronic health record system optimization: Systematic review of systematic literature reviews. **International journal of medical informatics**, Elsevier, v. 152, p. 104507, 2021.

NEUMANN, ME Peters M et al. Deep contextualized word representations. **arXiv preprint arXiv:1802.05365**, 2018.

NEWMAN-GRIFFIS, Denis et al. Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 28, n. 3, p. 516–532, 2021.

NOY, Natalya F et al. Bioportal: ontologies and integrated data resources at the click of a mouse. **Nucleic acids research**, Oxford University Press, v. 37, n. suppl_2, p. W170–W173, 2009.

OLIVEIRA, Lucas et al. Semclinbr – a multi institutional and multi specialty semantically annotated corpus for portuguese clinical nlp tasks. 01 2020.

OTTER, Daniel W; MEDINA, Julian R; KALITA, Jugal K. A survey of the usages of deep learning for natural language processing. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 32, n. 2, p. 604–624, 2020.

PAIM, Jairnilson et al. The brazilian health system: history, advances, and challenges. **The Lancet**, Elsevier, v. 377, n. 9779, p. 1778–1797, 2011.

PAIM, Jairnilson Silva. Thirty years of the unified health system (sus). **Ciência & Saúde Coletiva**, SciELO Brasil, v. 23, p. 1723–1728, 2018.

PALOJOKI, Sari et al. Semantic interoperability of electronic health records: Systematic review of alternative approaches for enhancing patient information availability. **JMIR medical informatics**, JMIR Publications Inc., Toronto, Canada, v. 12, n. 1, p. e53535, 2024.

PENG, Yifan; YAN, Shankai; LU, Zhiyong. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. **arXiv preprint arXiv:1906.05474**, 2019.

PERERA, Nadeesha; DEHMER, Matthias; EMMERT-STREIB, Frank. Named entity recognition and relation detection for biomedical information extraction. **Frontiers in cell and developmental biology**, Frontiers, p. 673, 2020.

PIRES, Telmo; SCHLINGER, Eva; GARRETTE, Dan. How multilingual is multilingual bert? **arXiv preprint arXiv:1906.01502**, 2019.

PODDER, V; LEW, V; GHASSEMZADEH, S. Soap notes.[updated 2021 sep 2]. **StatPearls [Internet]. StatPearls Publishing. Available from: https://www. ncbi. nlm. nih. gov/books/NBK482263**, 2021.

POMARES-QUIMBAYA, Alexandra; KREUZTHALER, Markus; SCHULZ, Stefan. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. **BMC medical research methodology**, Springer, v. 19, p. 1–20, 2019.

RADFORD, Alec et al. Improving language understanding by generative pre-training. 2018.

RAGHUPATHI, Wullianallur; RAGHUPATHI, Viju. Big data analytics in healthcare: promise and potential. **Health information science and systems**, Springer, v. 2, p. 1–10, 2014.

RAZA, Shaina; SCHWARTZ, Brian. Entity and relation extraction from clinical case reports of covid-19: a natural language processing approach. **BMC Medical Informatics and Decision Making**, Springer, v. 23, n. 1, p. 20, 2023.

RIVERA-ZAVALA, Renzo M; MARTÍNEZ, Paloma. Analyzing transfer learning impact in biomedical cross-lingual named entity recognition and normalization. **BMC bioinformatics**, BioMed Central, v. 22, n. 1, p. 1–23, 2021.

ROCHA, Naila Camila Da et al. Natural language processing to extract information from portuguese-language medical records. **Data**, MDPI, v. 8, n. 1, p. 11, 2022.

ROEHRS, Alex et al. Personal health records: a systematic literature review. **Journal of medical Internet research**, JMIR Publications Inc., Toronto, Canada, v. 19, n. 1, p. e13, 2017.

_____. Toward a model for personal health record interoperability. **IEEE journal of biomedical and health informatics**, IEEE, v. 23, n. 2, p. 867–873, 2018.

_____. Analyzing the performance of a blockchain-based personal health record implementation. **Journal of Biomedical Informatics**, v. 92, p. 103140, 2019. ISSN 1532-0464.

RUIZ-MARTÍNEZ, Juana María et al. Ontology learning from biomedical natural language documents using umls. **Expert Systems with Applications**, Elsevier, v. 38, n. 10, p. 12365–12378, 2011.

SÁEZ, Carlos et al. An hl7-cda wrapper for facilitating semantic interoperability to rule-based clinical decision support systems. **Computer methods and programs in biomedicine**, Elsevier, v. 109, n. 3, p. 239–249, 2013.

SALOMI, Maria Jose Amaral; CLARO, Priscila Borin et al. Adopting healthcare information exchange among organizations, regions, and hospital systems toward quality, sustainability, and effectiveness. **Technology and Investment**, Scientific Research Publishing, v. 11, n. 03, p. 58, 2020.

SARGSYAN, Astghik et al. The covid-19 ontology. **Bioinformatics**, Oxford University Press, v. 36, n. 24, p. 5703–5705, 2020.

SARIPALLE, Rishi Kanth. Fast health interoperability resources (fhir): current status in the healthcare system. **International Journal of E-Health and Medical Communications (IJEHMC)**, IGI Global, v. 10, n. 1, p. 76–93, 2019.

SCHNEIDER, Elisa Terumi Rubel et al. Biobertpt-a portuguese neural language model for clinical named entity recognition. In: **Proceedings of the 3rd Clinical Natural Language Processing Workshop**. [S.l.: s.n.], 2020. p. 65–72.

SHARMA, Sumit; JAIN, Sarika. Covido: an ontology for covid-19 metadata. **The Journal of Supercomputing**, Springer, v. 80, n. 1, p. 1238–1267, 2024.

SHEN, Yi-Cheng; HSIA, Te-Chun; HSU, Ching-Hsien. Analysis of electronic health records based on deep learning with natural language processing. **Arabian Journal for Science and Engineering**, Springer, p. 1–11, 2021.

SHETH, Amit P. Changing focus on interoperability in information systems: from system, syntax, structure to semantics. In: **Interoperating geographic information systems**. [S.l.]: Springer, 1999. p. 5–29.

SHULL, Jessica Germaine. Digital health and the state of interoperable electronic health records. **JMIR medical informatics**, JMIR Publications Inc., Toronto, Canada, v. 7, n. 4, p. e12712, 2019.

SILVA, Carolina Giordani da et al. Snomed-ct as a standardized language system model for nursing: an integrative review. **Revista Gaúcha de Enfermagem**, SciELO Brasil, v. 41, p. e20190281, 2020.

SILVERMAN, Greg M et al. Nlp methods for extraction of symptoms from unstructured data for use in prognostic covid-19 analytic models. **Journal of Artificial Intelligence Research**, v. 72, p. 429–474, 2021.

SIM, Jin-ah et al. Natural language processing with machine learning methods to analyze unstructured patient-reported outcomes derived from electronic health records: A systematic review. **Artificial intelligence in medicine**, Elsevier, p. 102701, 2023.

SINACI, A Anil; ERTURKMEN, Gokce B Laleci. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. **Journal of biomedical informatics**, Elsevier, v. 46, n. 5, p. 784–794, 2013.

SIVARETHINAMOHAN, R; SUJATHA, S; BISWAS, Pritha. Envisioning the potential of natural language processing (nlp) in health care management. In: IEEE. **2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic"(IEC)**. [S.l.], 2021. p. 189–193.

SLOANE, Elliot B; COOPER, Todd; SILVA, Ricardo. Mdira: Ieee, ihe, and fhir clinical device and information technology interoperability standards, bridging home to hospital to "hospital-in-home". In: IEEE. **SoutheastCon 2021**. [S.l.], 2021. p. 1–4.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. Bertimbau: pretrained bert models for brazilian portuguese. p. 403–417, 2020.

SOUZA, Fabio Capuano de. **BERTimbau: pretrained BERT models for brazilian portuguese**. 62 p. Master's Thesis (Dissertation in Electrical Engineering, in the Computer Engineering field.) — School of Electrical and Computer Engineering of the University of Campinas, Campinas, 2020.

SUN, Cong et al. Biomedical named entity recognition using bert in the machine reading comprehension framework. **Journal of Biomedical Informatics**, Elsevier, v. 118, p. 103799, 2021.

TANG, Paul C et al. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. **Journal of the American Medical Informatics Association**, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 13, n. 2, p. 121–126, 2006.

TAWEEL, Adel et al. Service and model-driven dynamic integration of health data. In: **Proceedings of the first international workshop on Managing interoperability and complexity in health systems**. [S.l.: s.n.], 2011. p. 11–17.

TINN, Robert et al. Fine-tuning large neural language models for biomedical natural language processing. **arXiv preprint arXiv:2112.07869**, 2021.

UROVI, Visara et al. An agent coordination framework for ihe based cross-community health record exchange. In: **VII Workshop on Agents Applied in Health Care, A2HC 2012**. [S.l.: s.n.], 2012. p. 29.

VASHISHTH, Shikhar et al. Reside: Improving distantly-supervised neural relation extraction using side information. **arXiv preprint arXiv:1812.04361**, 2018.

VRETINARIS, Alina et al. Medical entity disambiguation using graph neural networks. In: **Proceedings of the 2021 International Conference on Management of Data**. [S.l.: s.n.], 2021. p. 2310–2318.

WANG, Li et al. Archetype relational mapping-a practical openehr persistence solution. **BMC medical informatics and decision making**, BioMed Central, v. 15, n. 1, p. 1–18, 2015.

WANG, Samuel J et al. A cost-benefit analysis of electronic medical records in primary care. **The American journal of medicine**, Elsevier, v. 114, n. 5, p. 397–403, 2003.

WANG, Tiankai; DOLEZEL, Diane. Usability of web-based personal health records: an analysis of consumers' perspectives. **Perspectives in Health Information Management**, American Health Information Management Association, v. 13, n. Spring, 2016.

WOODS, Leanna et al. Show me the money: how do we justify spending health care dollars on digital health? **The Medical Journal of Australia**, Wiley, v. 218, n. 2, p. 53, 2023.

WRIGHT, Adam et al. Bringing science to medicine: an interview with larry weed, inventor of the problem-oriented medical record. **Journal of the American Medical Informatics Association**, BMJ Publishing Group BMA House, Tavistock Square, London, WC1H 9JR, v. 21, n. 6, p. 964–968, 2014.

XIAO, Cao; CHOI, Edward; SUN, Jimeng. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 25, n. 10, p. 1419–1428, 2018.

YANG, Lin et al. Discovering clinical information models online to promote interoperability of electronic health records: a feasibility study of openehr. **Journal of medical Internet research**, JMIR Publications Inc., Toronto, Canada, v. 21, n. 5, p. e13504, 2019.

YANG, Xi et al. Clinical concept extraction using transformers. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 27, n. 12, p. 1935–1942, 2020.

YUKSEL, Mustafa et al. An interoperability platform enabling reuse of electronic health records for signal verification studies. **BioMed research international**, Hindawi, v. 2016, 2016.

ZHA, Yue et al. Ontology attention layer for medical named entity recognition. **Applied Sciences**, MDPI, v. 14, n. 1, p. 421, 2024.

ZHANG, Tianyi et al. Revisiting few-sample bert fine-tuning. **arXiv preprint arXiv:2006.05987**, 2020.

ZHANG, Yijia et al. A hybrid model based on neural networks for biomedical relation extraction. **Journal of biomedical informatics**, Elsevier, v. 81, p. 83–92, 2018.

ZHOU, Jie et al. Graph neural networks: A review of methods and applications. **AI Open**, Elsevier, v. 1, p. 57–81, 2020.

ZUO, Zheming et al. Data anonymization for pervasive health care: Systematic literature mapping study. **JMIR medical informatics**, JMIR Publications Inc., Toronto, Canada, v. 9, n. 10, p. e29871, 2021.