Oncolog-IA : symbolic and numeric artificial intelligence for learning complexity of breast cancer cases and providing decision support for their therapeutic management

Akram Redjdal

Sorbonne Université

École doctorale Pierre Louis de santé publique

Laboratoire d'Informatique Médicale et
d'Ingénierie des Connaissances en e-Santé

ONCOLOG-IA

Utilisation des méthodes d'intelligence artificielle pour la détection automatique des cas complexes de cancer du sein et l'aide à la décision pour leur prise en charge thérapeutique

Akram REDJDAL

Thèse de doctorat en Informatique Biomédicale

Dirigée par Brigitte SEROUSSI

Présentée et soutenue publiquement le 28 Septembre 2023

Devant un jury composé de :

Brigitte	SEROUSSI	PU-PH	Directrice
Joseph	GLIGOROV	PU-PH	Co-directeur
Olivier	BODENREIDER	MD, PhD	Rapporteur
Silvana	QUAGLINI	PU	Rapporteur
Jacques	BOUAUD	PhD	Co-encadrant
Sandra	BRINGAY	PU	Présidente
Marc	CUGGIA	PU-PH	Examinateur
Lina	SOUALMIA	PU	Examinatrice

À mon père, pour sa confiance et ses conseils avisés.
À ma mère pour son soutien inébranlable.
À mon frère, pour son appui chaleureux.
À Jennifer, pour son amour inégalé.

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude envers ma directrice de thèse Brigitte Seroussi et mon encadrant Jacques Bouaud. Leur accueil lors de mon stage de master a marqué le début de cette aventure, et leur confiance en moi pour ce projet de thèse a été un catalyseur déterminant. Leurs conseils et leur soutien tout au long de ce travail ont été inestimables.

Un grand merci à Marie-Christine Jaulent pour m'avoir accueilli au sein du laboratoire et pour m'avoir offert l'opportunité d'y rester en tant qu'ingénieur de développement pendant un an, en attendant de trouver un financement pour ma thèse.

Je souhaite exprimer ma reconnaissance envers Silvana Quaglini et Olivier Bodenreider pour avoir rapporté ma thèse, ainsi qu'à Lina Soualmia, Sandra Bringay et Marc Cuggia pour avoir accepté de faire partie du jury.

Je remercie particulièrement Joseph Gligorov pour avoir accepté de co-diriger cette thèse et pour son implication cruciale dans la partie clinique de mon travail. Un merci également à Coralie Prebet, Marc-Antoine Benderra, et aux cliniciens de la RCP de sénologie de l'hôpital Tenon pour avoir contribué malgré leur emploi du temps chargé.

Un remerciement spécial à Tram-Anh Huynh et Natallia Novikava, dont la contribution directe à mon travail a été précieuse. Leur dévouement et leurs travaux ont enrichi cette recherche et contribué à son avancement.

Je remercie aussi, Gilles Guezennec (†2020), bien que n'ayant pas pu voir les résultats finaux de ce travail, il a joué un rôle précieux en me formant lors de mes débuts au LIMICS. Merci également à Sylvie Després, Xavier Tannier et Jean Charlet ainsi que l'ensemble des enseignants et chercheurs du laboratoire pour leur confiance, conseils et les discussions enrichissantes.

Un grand merci également à tous les doctorants du LIMICS, en particulier à mes "jumeaux de thèse" Chyrine Tahri et Jacques Hilbey, qui ont été mes confidents et camarades de route tout au long de cette aventure. Merci aussi à Manon Chossegros et Morgan Vaterkowski pour le partage du bureau et des moments du quotidien. Jacques Hilbey (avec Jean Charlet) méritent une mention spéciale pour avoir fourni un template qui a grandement facilité la rédaction de ce manuscrit.

Je souhaite, avant de finir, exprimer ma reconnaissance envers Isabelle Verdier pour son accompagnement précieux dans les procédures administratives parfois complexes.

Enfin un merci infini à mes parents, mon frère et à Jennifer pour leur amour, leur soutien indéfectible et leur compréhension tout au long de ce parcours académique. Leur encouragement a été ma source d'inspiration constante.

Résumé : De nombreux pays ont instauré l'organisation de réunions de concertation pluridisciplinaire (RCP) afin de promouvoir la décision collective des différents professionnels de santé impliqués dans la prise en charge des patients atteints de cancer. Cependant, l'impact des RCP sur la qualité des soins a été remis en cause car le bon fonctionnement des RCP est entravé par le manque de temps, la quantité des informations à gérer, et la complexité des cas discutés. Par ailleurs, les systèmes d'aide à la décision médicale (SADM), ont le potentiel d'améliorer la qualité des décisions de prise en charge de cancer du sein, mais ils sont encore très peu utilisés en routine, notamment parce qu'ils ne sont pas en adéquation avec les attentes des cliniciens qui les utilisent. Oncolog-IA est un projet de recherche, qui vise à utiliser des méthodes d'intelligence artificielle numériques et symboliques pour l'apprentissage des cas complexes de cancer du sein à partir d'un corpus de documents incluant les fiches issues des RCP extraites de l'EDS de l'AP-HP. Les fiches RCP sont préalablement structurées par la mise en œuvre de techniques de traitement du langage naturel. Une fois l'apprentissage de la complexité établi, l'objectif du projet est de proposer deux SADM selon la complexité des cas cliniques de cancer du sein :

1. Un système basé sur les guides de bonnes pratiques pour les cas non complexes
2. Un système basé sur un raisonnement par analogie pour les cas complexes, à travers le rappel des décisions prises pour des cas similaires.

Mots-clés : Intelligence artificielle, cancer du sein, traitement du langage naturel, systèmes d'aide à la décision.

Title of the thesis in English

Oncolog-ia: symbolic and numeric artificial intelligence for learning the complexity of breast cancer cases and providing decision support for their therapeutic management

Abstract: Many countries have introduced multidisciplinary tumor boards (MTBs) to promote collective decision-making by the various health professionals involved in the management of cancer patients. However, the impact of MTBs on the quality of care has been questioned because the proper functioning of MTBs is hampered by the lack of time, the amount of information to be managed, and the complexity of cases discussed. On the other hand, clinical decision support systems (CDSSs) have the potential to improve the quality of breast cancer management decisions, but they are still not used in clinical routine, notably because they are not in line with the expectations of the clinicians who use them. Oncolog-IA is a research project, which aims at using numerical and symbolic artificial intelligence methods for learning complex breast cancer cases from a corpus of documents including breast cancer patient summaries (BCPSs) extracted from the data warehouse of AP-HP hospitals. BCPS contents have been structured by implementing various natural language processing techniques, and algorithms were then trained to automatically detect the complexity of breast cancer cases. Once the complexity has been learned, the second objective of the project was to propose two decision support systems according to complexity:

1. A guideline-based decision support system for non-complex cases
2. A case-based decision support system for complex cases, through the recall of decisions taken for similar cases.

Keywords: Artificial intelligence, breast cancer, natural language processing, clinical decision support systems.

Contents

CHAPTER 1

Introduction

This chapter serves as the introductory gateway to a comprehensive exploration of clinical decision support systems in the context of breast cancer management. It begins by presenting various interconnected challenges within this domain, including the increasing incidence of breast cancer, the role of multidisciplinary tumor boards, the potential of clinical decision support systems, and the intricacies of managing complex clinical cases. As the chapter unfolds, the reader is guided through the transition from problem delineation to the formulation of research questions, illustrating how these seemingly distinct challenges are intrinsically linked.

1.1 Background

1.1.1 General context

Having replaced lung cancer as the most commonly diagnosed cancer globally, breast cancer is a significant health concern for women. With an estimated 2.3 million new cases diagnosed worldwide, it was by far the most commonly diagnosed cancer in women in 2020 (Sung *et al.*, 2021). In that same year, breast cancer took the lives of approximately 685,000 women, representing a significant proportion of cancer deaths among women, with 1 in 6 affected. By 2040, the number of newly diagnosed breast cancer cases is expected to increase by more than 40%, with approximately 3 million cases being diagnosed annually. Even more worrying is the fact that deaths from breast cancer are predicted to rise by over 50%, from 685,000 in 2020 to 1 million in 2040 (Arnold *et al.*, 2022). These projections are primarily due to population growth and aging, and changes in incidence rates may further impact these numbers.

In many countries, multidisciplinary tumor boards (MTBs) have been introduced to promote the collective decision of health professionals involved in managing breast cancer patients (Muggia, 1984). Medical professionals from various specialties gather to discuss the best possible treatment plan for a patient with cancer. The board typically includes medical oncologists, surgeons, radiation oncologists, pathologists, radiologists, and other healthcare providers as needed (like psychologists and nutritionists). During a tumor board meeting, the medical team reviews the

patient's medical history, reports, imaging studies, and other relevant information to make an accurate diagnosis and determine the best course of treatment. The team will discuss all available treatment options, including surgery, radiation therapy, chemotherapy, hormone therapy, and immunotherapy, and will consider the potential benefits and risks of each.

One of the key benefits of a multidisciplinary tumor board is that it allows for collaboration and communication among different medical specialties. By working together, the team can develop a more comprehensive and effective treatment plan that considers each patient's unique needs and circumstances while integrating state-of-the-art and clinical practice guidelines (CPGs).

However, evidence around the effectiveness of MTBs on cancer care has been actively analyzed. While studies have shown that MTBs are effective in improving the compliance of therapeutic decisions to CPG recommendations (Kesson *et al.*, 2012; van Hoeve *et al.*, 2014; Brar *et al.*, 2014), their benefits are being challenged (Keating *et al.*, 2013). Indeed, clinical teams can be affected by staff shortages, workload, the continuously increasing number of cases to discuss, and disciplinary diversity. So the impact of MTBs on the quality of care has been questioned (Soukup *et al.*, 2022; Blayney, 2013; El Saghir *et al.*, 2013).

Overall, MTBs are an important part of modern cancer care, and they play a critical role in ensuring that patients receive the best possible treatment outcomes. However, the organization of MTBs should be improved to guarantee all cancer patients receive the best treatment and create the best possible environment for clinicians to collaborate and make informed decisions. By doing so, we can help guarantee that every patient receives the highest standard of care and the best chance for successful treatment. In France, one of the objectives of the first 2003 cancer plan was to ensure that 100% of new cancer patients would benefit from a discussion meeting concerning their case. The organization of the MTBs is defined in article D. 6124-131 of the French Public Health Code.

1.1.2 Clinical decision support systems

Clinical Decision Support Systems (CDSSs) are important tools for modern healthcare, providing computer-based assistance to clinicians in making decisions for their patients. As we will further see in detail in section 2.1, CDSSs have been recognized for improving physicians in making personalized treatments for cancer patients (Hammond *et al.*, 1994). By providing evidence-based information according to patient-specific data, CDSSs can help clinicians identify high-risk patients, refine diagnoses, recommend appropriate treatment options, and monitor treatment progress.

In the paradigm of evidence-based medicine (Evidence-Based Medicine Working Group, 1992), numerous CDSSs have been developed and evaluated to promote evidence-based clinical decision-making in oncology (see section 2.1.1). Such systems often rely on the knowledge contained in CPGs which summarize the state of the art. However implementing guideline-based CDSSs in clinical practice faces technical challenges, including semantic interoperability with EHR systems, inconsistent EHR storage, and the need for data validation. Maintenance of guideline-based CDSSs and managing multiple guidelines present further complexities. Non-adherence to CDSS recommendations can occur in uncommon clinical scenarios with limited scientific evidence (Voigt & Trautwein, 2023).

A few years ago, a new paradigm emerged in the medical domain. Precision medicine, sometimes known as "personalized medicine", is an innovative approach to tailoring disease prevention and treatment taking into account differences in people's genes, environments, and lifestyles. The goal of precision medicine is to target the right treatments to the right patients at the right time (Gameiro *et al.*, 2018). As it will be explained in section 2.1.2, recent developments in Artificial

Intelligence (AI) have exhibited great promise in revolutionizing the field of clinical oncology by effectively tackling multiple critical aspects throughout the entire journey of cancer care (Corti *et al.*, 2023).

Many CDSSs based on machine and deep learning have been developed. These systems do not rely on explicit knowledge but on regular patterns discovered in past data. Huge amounts of (available) clinical data are required for training and reuse in new and similar situations. But even if AI holds tremendous potential in clinical oncology, there are key challenges that must be addressed to successfully integrate AI into routine care.

Recent research (Norgeot *et al.*, 2020; Thompson *et al.*, 2018) highlighted the limited number of prospective trials and randomized clinical trials for deep learning models, indicating the need for further validation and evidence. Challenges such as data limitations, model interpretability, and ensuring clinical validity, utility, and usability of AI models must also be overcome. Transparency and the need for trained clinicians in AI-based CDSSs pose additional hurdles.

Overall, despite their potential benefits, the use of CDSSs whether knowledge-based or non-knowledge-based in clinical routine remains limited (Beauchemin *et al.*, 2019). There is still significant work to be done in this field to improve and expand their implementation. ①

1.1.3 Complexity of clinical cases

DESIREE is a European project (Bouaud *et al.*, 2020b), that focused on enhancing care for primary breast cancer patients through a cutting-edge web-based platform. One of its notable features is the guideline-based decision support system (GL-DSS), extensively presented in section 2.2.4.1. As a part of the DESIREE initiative, we conducted an evaluation of the GL-DSS, the knowledge base of which was based on French guidelines published in 2016.

Through this evaluation, we discovered instances where the system did not generate therapeutic proposals for certain patient cases or recommended treatments that were not followed by MTB clinicians. We discovered that these cases were not covered by the CPGs or had peculiar characteristics, and therefore required in-depth multidisciplinary discussions during the MTBs. After discussing with oncology clinicians regarding these profiles, experts expressed that these types of clinical cases pose challenges. We termed such clinical profiles **complex cases.** Consequently, they favor an alternative form of decision support, for instance, clinicians said that for complex cases, the recall of similar cases along with the MTB decisions made for them would be a good support for determining the appropriate care plan. ②

Patient clinical cases may be of various levels of complexity, and more time should be given to MTBs to discuss complex cases and avoid fatigue impairing good decision-making (Soukup *et al.*, 2019); however, there is no a priori definition of breast cancer complexity and very few tools are available that assess cancer complexity (Soukup *et al.*, 2020).

1.1.4 Clinical notes

Hospital clinical documents, such as discharge summaries and clinical notes, are a valuable source of information for a variety of purposes. It was estimated that 80% of hospital data are collected in the form of text (Raghavan *et al.*, 2014). However, the free text format of these documents can make it difficult to extract and process the contained information in a structured way. This can limit the usefulness of such information for clinical care, research, and other applications.

One way to address this challenge is to use information extraction (IE) techniques to automatically structure the text of clinical documents. IE involves identifying and extracting specific pieces

of information from text, such as patient demographics, diagnoses, or procedures. This structured data can then be used for a variety of purposes, such as clinical decision support, research, or medical coding.
The use of IE for clinical documents is a growing body of research in medical informatics. As IE techniques continue to improve, they have the potential to have a significant impact on the way that clinical data is used.

Regarding cancer care, during the MTBs, clinicians refer to a document, usually produced by the physician in charge of the patient before the MTB. The physician gathers all the information needed to make a decision for his patient, including clinical history, radiology results, pathology results, response to treatment, etc., and summarizes all the information in a textual document. This document is shared among MTB participants and will be completed by the MTB decision. this document is called the Breast Cancer Patient Summary (BCPS).

BCPSs provide a portrait of patients with all the relevant information that MTB clinicians need to know to make the best patient-specific therapeutic decision. It is a crucial document for the MTB. However, this document is written in natural language, there are many abbreviations and specialized terms depending on the health professional specialty of the BCPS's author. This makes the use of the content BCPSs far from being straightly processed by a CDSS. ③

1.2 Research questions

Overall, MTBs play a crucial role in cancer care by facilitating collaborative decision-making among healthcare professionals. However, despite their potential to improve MTBs, clinical decision support systems are not routinely used in the management of cancer patients ①. This raises the main research question: **How can a clinical decision support system be developed to effectively assist multidisciplinary tumor board clinicians in their decision-making process?**

To address this question, we investigated the acceptance of CDSSs ②, we noticed that clinicians often considered it useless to apply guideline-based decision support systems to complex cases, as these cases are usually not covered by guidelines. Hence, the research question arises: **What can be done to make clinicians accept guideline-based decision support systems by taking into account the complexity of clinical cases and how can we effectively update guideline-based CDSSs?** Additionally, **How can we assist clinicians in managing complex cases during the decision-making process?**

Finally, textual cancer summaries, such as BCPSs, serve as vital sources of information during MTBs ③. However, these documents are typically written in natural language format which poses challenges in efficiently utilizing the information within BCPSs for decision support. Consequently, another research question emerges: **Can we effectively create a system that takes BCPSs as input and provides personalized treatment recommendations for patients?** The aim is to capture the most important information available in the BCPSs and use it to provide decision support, ultimately saving time for MTB clinicians.

1.3 Objectives

The main hypothesis of this work can be summed up in one expression: "*One size does not fit all!*". We know from previous experiences that guideline-based systems have the potential to improve the compliance of MTBs decisions with the guidelines (Seroussi *et al.*, 2012a), we also know that these systems are limited when it comes to complex clinical cases (Redjdal *et al.*, 2021b).

Therefore, the goal of this thesis is to create a CDSS that supports MTB decision-making for the management of breast cancer patients. According to our assumption, it will first identify complex cases enabling MTB clinicians to optimizer MTB's sequencing and focus on these cases, which require deeper conversations. Then, depending on the case complexity, the appropriate therapeutic decision support tool will be triggered. Beyond recognizing complex cases, the complexity classifier will serve as a triage system to provide adequate decision support.

- **Support for complexity classification:** Since there is no definition of complexity, we proposed to learn complexity. An algorithm that classifies patient cases as complex or non-complex has been developed through machine learning using different feature extraction techniques (see Chapter 4). Beyond serving as a triage system to the main CDSS, it can support clinicians in the organization of the MTB by prioritizing discussions for complex patients.

- **Support for non-complex cases:** As these cases are simple, we assume they are correctly handled by the clinical practice guidelines. Therefore, for these cases, we reuse the guideline-based system developed within the DESIREE project, as documented in (Bouaud *et al.*, 2020a), but we have to adapt it to more recent breast cancer CPGs (see Chapter 5).

- **Support for complex cases:** As these cases are not adequately covered by CPGs, a case-based decision support approach is proposed. Chapter 6 presents a methodology to detect the most similar patients to a given patient. This system will recommend treatment options to MTB clinicians based on the decisions made for similar patients, thereby reproducing the medical reasoning used by clinicians during their decision-making process.

An important part of this work is the processing of textual BCPSs using natural language processing (NLP) techniques. Indeed, a mandatory preliminary task is to transform BCPS contents into a formal structured data format that enables the use of both guideline-based and case-based CDSSs, as well as the complexity determination task. This processing will extract relevant information from BCPSs, allowing the system to provide comprehensive decision support based on patient clinical situations.

By accomplishing these objectives, we aim to make a contribution to the decision support field by the development of a CDSS that enhances the decision-making process within multidisciplinary tumor boards. Through the integration of guideline-based and case-based decision support, as well as the utilization of NLP techniques, this project aims to provide a robust and effective CDSS that addresses the diverse needs of MTB clinicians.

1.4 Outline

The manuscript is organized as follows. Chapter 2, exposes the state-of-the-art in each of the main fields we have mobilized to carry out this research. The next four chapters describe the work we have done. Each of these chapters is written as an article and follows the IMRaD structure. Figure 1.1 provides a graphical outline for these chapters, illustrating their inputs, and their outputs. It can be summarized as follows :

Chapter 3 explains the transformation of textual BCPSs into a structured data format by combining rule-based and machine-learning NLP methods.

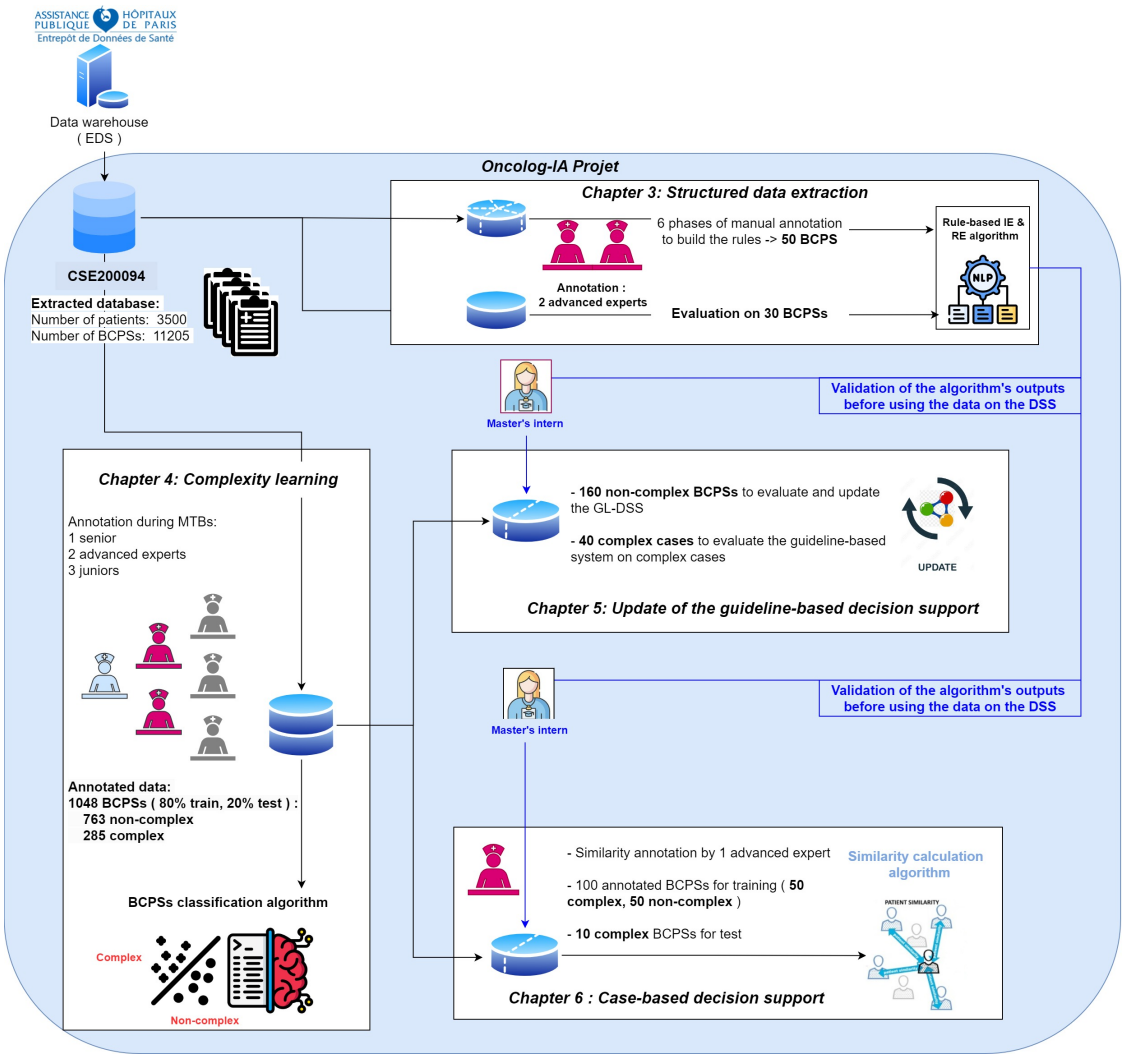


Figure 1.1: Outline of the project. *BCPS: Breast Cancer Patient Summary; IE: Information Extraction; RE: Relation Extraction; GL-DSS: Guideline-based decision support system of the DESIREE project*

Chapter 4 deals with breast cancer case complexity. We compare state-of-the-art methods to automatically classify patient cases as complex or non-complex from BCPs using supervised machine learning.

Chapter 5 describes the mapping of the structured data extracted in Chapter 3 into the breast cancer knowledge model to feed the GL-DSS of DESIREE, and the update to the knowledge base of the system to provide recommendations according to the most recent guidelines.

Chapter 6 tackles the problem of patient similarity. We implement a methodology based on state-of-the-art similarity metrics to retrieve similar patients for a given patient, providing a case-based CDSS.

Finally the last chapter concludes the manuscript.

1.5 Data exploitation

1.5.1 Hospital data warehouse

Assistance Publique-Hôpitaux de Paris (AP-HP), also known as Greater Paris University Hospitals, stands as the largest public university hospital network in France, boasting 39 distinct sites. Apart from its primary roles in patient care, education, and research, AP-HP plays a pivotal role in collecting and preserving patient data, with due consent, within a specialized clinical data warehouse named "Entrepôt de Données de Santé" (EDS). EDS is purposefully designed to facilitate healthcare research and innovation. It offers researchers access to a vast and diverse dataset, enabling studies on various health-related topics and supporting the development of cutting-edge healthcare applications, including decision-support tools and personalized medicine. Key features of EDS include:

- Storage of extensive clinical data, encompassing patient demographics, diagnoses, procedures, medications, and laboratory results.
- Secure web portal access for researchers, supplemented with computational tools.
- Regular updates with fresh data from AP-HP.

1.5.2 Data extraction

In our work, the aim is to utilize advanced technologies such as machine learning and deep learning associated with symbolic AI methods to develop a clinical decision support system that assists MTB clinicians in their decision-making process for the therapeutic management of breast cancer patients. To ensure the validity and ethical approval of the project, it has been reviewed and validated by the institutional review board at AP-HP (CSE200094). For our research, we were granted access to the breast cancer patient summaries of patients diagnosed with primary breast cancer between 2018 and 2022 and treated at Tenon Hospital, which is part of AP-HP. These BCPSs are accessible within EDS. As a result of the extraction, we had access to a database consisting of 3,500 patients diagnosed with breast cancer. Among them, we had access to 11,205 BCPSs, with each patient discussed in at least one MTB. In the next subsection, we will describe in detail how we used this database in the different chapters.

1.5.3 Datasets in the thesis work

The database extracted for the project **CSE200094** was used to extract datasets that served as training and evaluation for each of the different tasks of the project, it can be summarized as follow :

- For the structured data extraction (SDE) task (chapter 3), we performed a random selection without duplication, yielding a dataset of **80 BCPSs**. Among them, we used 50 BCPSs to develop rules for SDE. This dataset is denoted as the *SDE learning dataset*. Then, we used the rest of the BCPSs (30 BCPSs) to evaluate the rule-based SDE algorithm. This dataset was pre-annotated using a preliminary version of the algorithm and manually corrected by an advanced expert. It is referred to as the *SDE evaluation dataset*.
- In parallel, we selected **1,048 BCPSs** representing patients discussed during MTBs of Tenon Hospital between November for the complexity learning task (Chapter 4). A panel of experts,

comprising one senior, two advanced experts, and three juniors, annotated these BCPSs as either complex or non-complex. This corpus served as for machine learning algorithms designed for complexity detection and was named the *complexity learning dataset*. Of this dataset, 80% was allocated for training, while the remaining 20% was used for testing purposes.

- Using the complexity annotations, we further derived **160 non-complex cases** from the *complexity learning dataset*. This subdataset was used in evaluating and updating the GL-DSS as detailed in Chapter 5. it was named the *non-complex subdataset*.

- Finally, for the similarity calculation task (Chapter 6), we also selected two subdatasets from the *complexity learning dataset*. The first subdataset comprised **100 BCPSs** representing patients in the same clinical situation (patients who underwent surgery without neoadjuvant treatment, further referred to as patients in scenario D). Among these, 50 were previously classified as complex and 50 as non-complex. An advanced expert grouped this dataset into clusters of similar patients. It was named the *similarity learning dataset* and used to train algorithms for similarity calculation. Additionally, a final dataset consisting of **10 complex cases in scenario D** was employed to evaluate the similarity calculation algorithm. An advanced expert calculated, for each of the 10 BCPSs, the top 5 most similar patients from the *similarity learning dataset*. This dataset was designated as the *similarity evaluation dataset*.

For a visual representation of the dataset organization throughout the thesis, please refer to Figure 1.2.

1.6 Published works

In our research journey, we have published a preliminary paper (Redjdal *et al.*, 2021c) that introduced the concept of complex cases, laying the foundation for our exploration in this area. Additionally, we have contributed to the publication of two scoping reviews that will be discussed in detail in Chapter 2 (Novikava *et al.*, 2023; Azarpira *et al.*, 2022).

During this research, we explored the use of semantic annotators for processing BCPSs. This led to a publication (Redjdal *et al.*, 2022a) highlighting the importance of going beyond relying solely on semantic annotators for comprehensive structured data extraction from these textual documents. This publication served as the inspiration for chapter 3 of our work, where we describe the development of an NLP pipeline for structured data extraction from BCPSs.

Methods and results obtained in Chapter 4 have been published in two papers (Redjdal *et al.*, 2022b, 2021a). However, the findings and contributions from Chapter 5 and Chapter 6 have not been published yet.

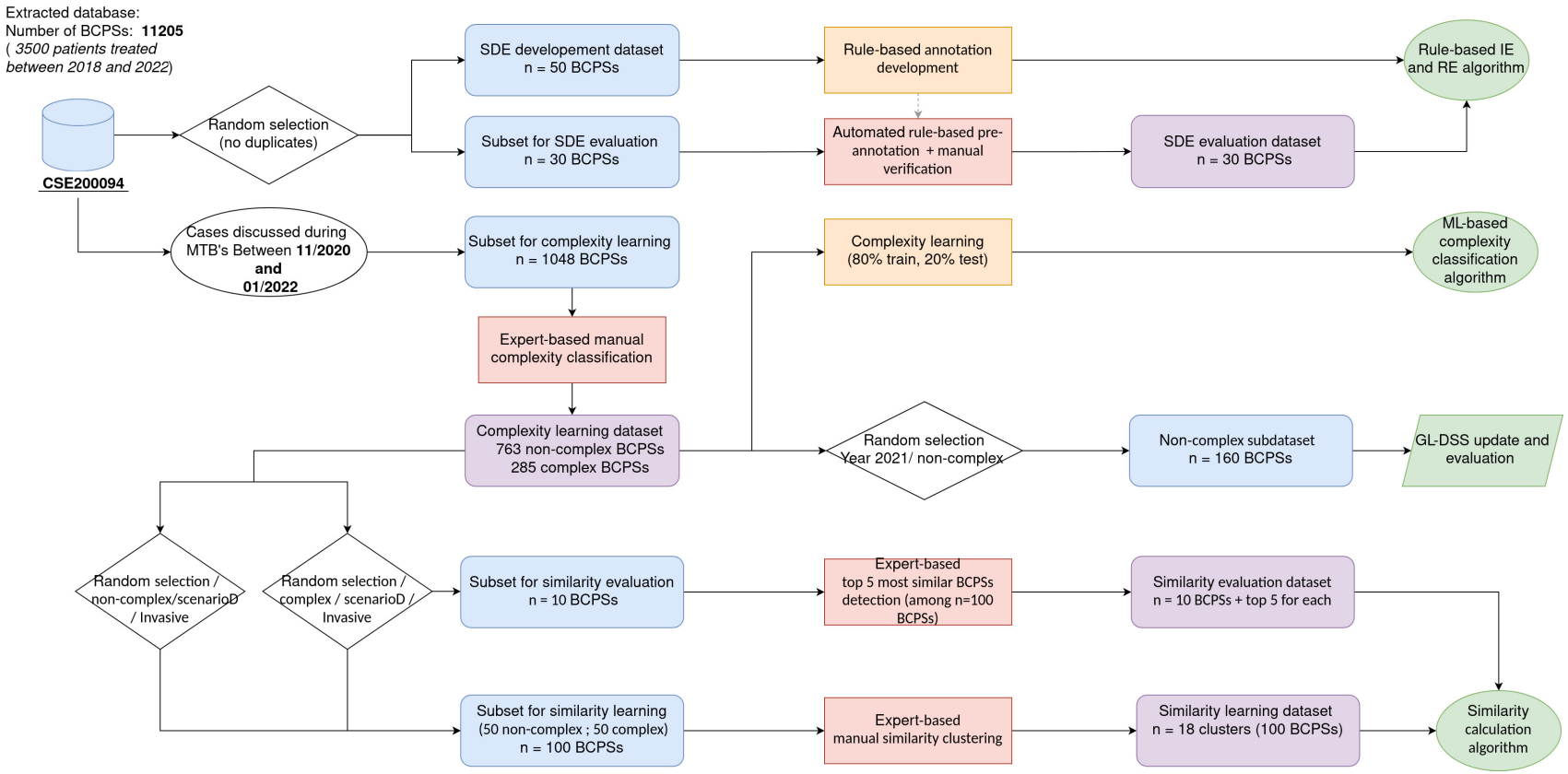


Figure 1.2: Consort-like Diagram Illustrating Dataset Utilization. *BCPS: Breast Cancer Patient Summary; IE: Information Extraction; RE: Relation Extraction; SDE: Structured Data Extraction; ML: Machine Learning; GL-DSS: Guideline-based decision support system of the DESIREE project; Scenario D: Patients who have undergone surgery without neo-adjuvant treatment*

CHAPTER 2

State-of-the-art

In this work, we aim at alleviating the burden of overloaded agendas of breast cancer multidisciplinary team meetings by providing clinicians with a multifaceted computerized decision support tool. The projected tool should first operate as a triage system (Fernandes et al.*, 2020) to categorize patient cases as complex cases deserving particular attention and discussion of the meeting participants, or non-complex cases for which standardized management based on guidelines could be recommended. In both situations, dedicated decision support paradigms can be used. In this application context, most, if not all, information about patient cases is provided as unstructured data, in text form. So while we are working on decision support, a main aspect of this project is the use of natural language processing methods to select and extract relevant information from the texts and turn them into structured coded data for further processing, here decision support. That is why this chapter is divided into two main parts. In the first section, we review the existing systems and approaches to decision support for the treatment of cancer, with a focus on breast cancer (sections 2.1,2.2 and 2.3). In the next section (section 2.4), we focus on natural language processing techniques, presenting state-of-the-art methods for named entity recognition, relation extraction, and text classification.*

Regarding the clinical decision support systems, our review mainly concerns systems for therapeutic management, as it is the main focus of this thesis, not diagnosis decision support.

2.1 Clinical decision support systems for therapeutic management in oncology

Medical decision-making is a critical process in which healthcare professionals continuously evaluate and make choices regarding patient care Bonatti *et al.* (2009). This complex and dynamic task involves considering various factors, including patient history, symptoms, test results, and available treatment options.

To support and enhance medical decision-making, CDSSs have emerged as valuable tools in the healthcare domain (Agharezaei *et al.*, 2013). CDSSs leverage advanced technologies, such as artificial intelligence and knowledge management, to assist healthcare professionals in their decision-making processes (Parshutin & Kirshners, 2013). By utilizing CDSSs, healthcare professionals can benefit from improved efficiency and better patient outcomes. CDSSs assist in reducing diagnostic errors, identifying potential drug interactions, suggesting appropriate treatment options, and ensuring adherence to clinical practice guidelines. Figure 2.1 from Wang *et al.* (2023) illustrates the role of CDSSs and their application in the clinical workflow :

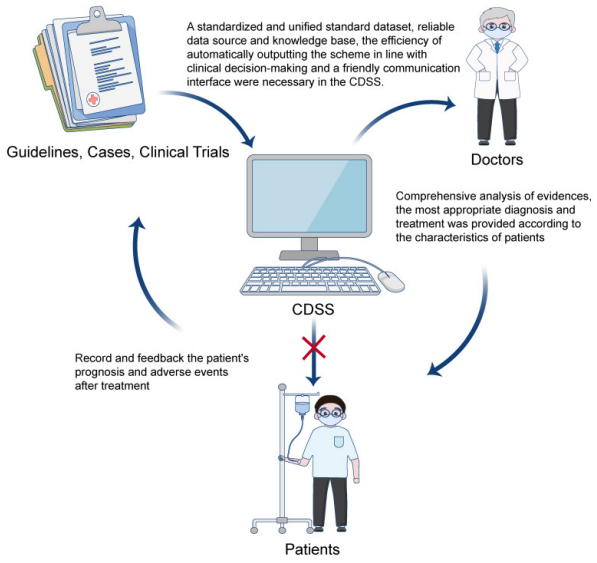


Figure 2.1: Applications of CDSS in clinical work (Wang *et al.*, 2023)

According to Greenes (2014), such systems can be categorized into two distinct groups belonging to two different artificial intelligence approaches. The first group is composed of knowledge-based CDSSs, which function as expert systems. These systems consist of a knowledge base, an inference engine, and a communication mechanism to obtain patient data. The knowledge base contains rules, in the form of "if-then" rules. The inference engine synthesizes these rules with patient data to generate new data and recommendations.

The second group is made of data-based (i.e. non-knowledge-based) CDSSs, which employ artificial intelligence techniques belonging to the machine learning field. These systems learn from past experiences and detect patterns from data using approaches such as neural networks or genetic algorithms. Non-knowledge-based CDSSs have the advantage of learning from complex data relationships, while knowledge-based CDSSs provide explanations for their recommendations based on explicit rules.

2.1.1 Knowledge-based CDSSs

There is a long history of knowledge-based DSSs in oncology (Leaning *et al.*, 1992) since the pioneering ONCOCIN system, a chemotherapy protocol advisor (Kent *et al.*, 1985)(Shortliffe, 1986). These systems were expert systems, so-called since the knowledge they relied on to deliver medical advice was often the local expertise of the system developers. One drawback of this approach was the difficulty to export these tools to other clinical settings for both technical issues (non-interoperability) and medical issues (content disagreement for non-shared knowledge). Accompanying the development of the evidence-based medicine (EBM) paradigm, many CDSSs have then been developed based on clinical practice guidelines (CPGs) the contents of which are expected to be shared by the medical community (Haines & Jones, 1994)(Gordon, 1996). Moreover, CPGs provide an immediate endpoint to assess the efficiency of care, available far before the ultimate endpoint of survival (Lobach & Hammond, 1997).

In oncology, numerous CDSSs based on CPGs have been developed and evaluated, and have been shown to support evidence-based clinical decision-making (Klarenbeek *et al.*, 2020b).

In the realm of cancer pain management, Bertsche *et al.* (2009) introduced the AiDPainCare system, which successfully reduced the number of patients with guideline deviations, ensuring optimal pain control while adhering to established guidelines. Verberne *et al.* (2012) evaluated the CEAwatch tool for colorectal cancer, which effectively decreased the workload for follow-up carcino-embryonic antigen (CEA) testing, demonstrating high adherence to the follow-up scheme. Magrath *et al.* (2018) showcased the impact of CDSSs on guideline adherence in colon cancer, reporting significant increases in adherence. Adeboyeje *et al.* (2017) investigated the use of the CSF DS tool in lung cancer management, reporting a substantial decrease in the use of Colony-stimulating factors (CSFs) when the decision-making was assisted by the tool. Ciprut *et al.* (2020) demonstrated the effectiveness of the CROC tool in reducing the risk of inappropriate imaging for prostate cancer. Recently, Lanzola *et al.* (2023) developed a system that provides cancer patients with coaching advice and supports their clinicians with suitable decisions based on clinical guidelines.

However, the successful implementation of guideline-based CDSSs in clinical practice faces several challenges. Most importantly, maintenance of guideline-based CDSSs can be challenging, guidelines, pathways, and workflow models need to be updated to reflect the latest evidence. The lack of interoperability and automatic update capabilities in the knowledge base of guideline-based CDSSs complicates these tasks (Cánovas-Segura *et al.*, 2019) (Fux *et al.*, 2020). The management of multiple guidelines and potential conflicts or inconsistencies in recommendations is another challenge in oncology CDSSs (Bilici *et al.*, 2018). Moreover, uncommon clinical profiles and situations with limited scientific evidence are reasons for nonadherence to guideline-based CDSS recommendations (Waks *et al.*, 2013; Bouaud *et al.*, 2012; Seroussi *et al.*, 2012b).

A recent paper by Voigt & Trautwein (2023) discusses "*the beneficial effects of CDSS on guideline adherence as well as technical and structural requirements for CDSS implementation in clinical routine*".

2.1.2 Non-Knowledge-based CDSSs

With the emergence of precision medicine, non-knowledge-based CDSSs offer an alternative approach to knowledge-based systems. They use machine learning techniques to enhance decision support. In the following subsection, we explore the application of non-knowledge-based CDSSs in oncology and their potential benefits. We focus on systems used for the treatment part of the

patient's pathway. To get an overview of the latest systems for screening, and diagnosis, the works mentioned by Tran *et al.* (2021) and Kann *et al.* (2021) provide details about methods used in all the patient's pathways from prevention to follow-up.

AI has shown promising potential in transforming clinical oncology by addressing various touchpoints along the cancer care pathway. For instance, AI algorithms have been developed to predict an individual's risk of developing cancer by leveraging diverse data sources such as genomics (Ming *et al.*, 2020), imaging (Bibault *et al.*, 2020), internet search history (White & Horvitz, 2017), and family history (Ming *et al.*, 2020). These algorithms enhance risk prediction beyond traditional models, enabling targeted screening and early interventions.

Risk stratification and prognosis have seen advancements through AI applications. Machine learning techniques integrating genomic (Scott *et al.*, 2017), imaging (Kann *et al.*, 2020), and clinical data (She *et al.*, 2020) have improved risk stratification models. AI algorithms have also been investigated for determining optimal initial treatment strategies by analyzing genomic (Scott *et al.*, 2017) and radiomic data (Sun *et al.*, 2018). Additionally, AI models that consider tumor mutational burden, copy number alteration, and microsatellite instability have shown promise in predicting response to immunotherapy (Xie *et al.*, 2020).

The assessment of treatment response has benefited from AI applications as well. Automated deep learning models have been developed to assess treatment response using criteria such as RANO assessment (Kickingereder *et al.*, 2019) and RECIST response for immunotherapy (Arbour *et al.*, 2021). RECIST and RANO are quantitave response assesment criteria.
Subsequent treatment strategies have been guided by AI algorithms considering factors such as prior treatments and restaging imaging (Xu *et al.*, 2019).

Follow-up care has also seen potential improvements through AI. Predicting recurrence risk and late toxicity based on radiomic features has shown promise in tailoring follow-up plans (Chang *et al.*, 2019). Additionally, AI leveraging EHR data has demonstrated the ability to triage patients for personalized, escalated follow-up strategies (Hong *et al.*, 2020). Other works regarding patients who experience relapse that cannot be effectively treated have been done. Machine learning has demonstrated promise in this context too, by assisting in identifying patients with a high risk of mortality and encouraging physicians to engage in meaningful discussions about their values, preferences, and available options to enhance their quality of life (Ramchandran *et al.*, 2013).

While AI shows great promise in clinical oncology, there are challenges to be addressed for successful translation into routine care. In fact ; a recent review by Nagendran *et al.* (2020) found only nine prospective trials for deep learning models in imaging and only two published randomized clinical trials.
These challenges include data limitations (Norgeot *et al.*, 2020; Thompson *et al.*, 2018), model interpretability (Doshi-Velez & Kim, 2017), and validation of clinical impact. Ensuring clinical validity, utility, and usability of AI models is crucial (Kang *et al.*, 2020; Kim *et al.*, 2019; Liu *et al.*, 2019a). Transparency and limited evidence can also contribute to unease regarding the usage of non knowledge-based CDSSs (Sutton *et al.*, 2020). Other aspects make the implementation of CDSS difficult, like the need of training clinical personnel involved in CDSS usage (Klarenbeek *et al.*, 2020a).

2.1.3 CDSS applied to breast cancer management

As we work on breast cancer, we published a scoping review on CDSSs, used for treatment of breast cancer (Novikava *et al.*, 2023). In this review we performed a literature search, using PubMed and Web Of Science, to retrieve papers published between 2000 and 2023, describing CDSSs applied to breast cancer management. We focused on articles about CDSSs that support treatment decisions. The following exclusion criteria were thus applied: (i) CDSSs focused on breast cancer screening; (ii) studies that use CDSSs to support all types of image analysis for diagnosis; (iii) studies that use CDSSs to support genetic analysis or biomarker discovery decisions, excluding treatment decisions; (iv) CDSSs applied to specific groups of patients (e.g. geriatric patients); (v) papers available only in the form of abstracts because of insufficient details. At the end we selected a total of 17 article that describe 15 different CDSSs for breast cancer management.

An analysis of the selected papers was done to categorize CDSSs according to their objective and check whether they were used or not in the clinical routine. The main categories we selected to classify the systems are presented below :

- **Risk calculators (RCs):** RCs are systems that are not knowledge-based, and use predictive modeling to provide a probability concerning the positive impact of a treatment on the survival rate or calculating the 5-year or lifetime risk of developing a new breast cancer. Among RCs for breast cancer, we found 6 systems that we describe below:

  - *Treatment Benefit Estimation* : Systems that estimate the benefit of a treatment on the survival rate, for breast cancer, we found **OncoAssist** by Jacob *et al.* (2019), **PREDICT** (Wishart *et al.*, 2010; Candido Dos Reis *et al.*, 2017) and **Adjuvant!Online** (Campbell *et al.*, 2009). These systems estimate the benefit of adjuvant treatment after breast cancer surgery. Adjuvant!Online focuses on adjuvant chemotherapy vs hormonotherapy while PREDICT and OncoAssist explore other treatment options.
  - *Recurrence Risk Assessment* : Systems that estimate the risk of recurrences of breast cancer. **CTS5 Calculator** (Dowsett *et al.*, 2018) calculates the risk of late distant recurrence (after 5 years of endocrine treatment). **RCB Calculator** (Sahoo *et al.*, 2022) calculates the residual cancer burden after neoadjuvant chemotherapy.
  - *Risk Estimation and Life Expectancy Reduction* : an example of these systems in breast cancer is **CancerMath** (Michaelson *et al.*, 2011), it estimates the risk of the reduction in life expectancy and survival rate.

- **Therapeutic decision support:** systems that provide a patient-specific care plan. We distinguish two main categories:

  Guideline-based decision support systems : We found 5 systems that align with established guidelines to assist in decision-making. Some of them follow an automatic approach to decision support like **Watson for Oncology** from IBM (Somashekhar *et al.*, 2018; Jie *et al.*, 2021) and **OncoGuide** (Hendriks *et al.*, 2019, 2020). These systems take patient data as input and the give the appropriate recommendations according to the patient case. Other systems like Oncodoc & Oncodoc2 (Seroussi *et al.*, 2007, 2001) propose a documentary approach to decision support, where we navigate through the decision tree of OncoDoc, by answering the questions of the system until we get the appropriate recommendation. Finally **OncoCure** focuses on guidelines to provide treatment plans after a breast cancer surgery (Eccher *et al.*, 2014).

Mixed systems: Systems that combine different methods including systems that automatically detect inclusion criteria to clinical trials. We consider here that the trial is the treatment proposed to the patient.
Among these systems, **CancerLinQ** developed by Potter *et al.* (2020) and **MATE** (Patkar *et al.*, 2012) combine a guideline-based system with an eligibility criteria identification module. Another system **(CLARIFY)** by Torrente *et al.* (2022) combines risk calculation algorithms eligibility criteria identification, and finally **DESIREE** combines guideline-based (Bouaud *et al.*, 2020b), case-based, and experience-based decision support to provide treatment recommendations (Seroussi *et al.*, 2018; Pelayo *et al.*, 2020).

We investigated whether these systems were employed or currently in use in regular clinical practice. Most risk calculators (Cancer Math, Residual Cancer Burden Calculator, OncoAssist, Adjuvant!Online, and PREDICT) have either been utilized or are currently being used in routine practice (Adjuvant!Online has not been updated and now recommends the use of PREDICT). These systems can be accessed online and function as standalone tools, requiring minimal data for operation. This makes it convenient for clinicians to employ them for risk assessment. In addition to that, recent french guidelines recommend to use PREDICT for specific types of patient profiles. This suggest that the information provided by these tools is important, and clinicians need them in their daily practice.
On the contrary, systems that provide treatment careplan have seen limited adoption in clinical routine. DESIREE and CLARIFY are not yet in widespread use, primarily due to interoperability challenges with electronic health records (EHRs). Similarly, systems like OncoDoc and its updated version OncoDoc2 were utilized in clinical practice for several years, resulting in increased adherence to guidelines (Seroussi *et al.*, 2007), but they are no longer in use due to technical constraints. Other systems that provide care plan recommendations, such as MATE, OncoCure, and OncoGuide, have limited usage in a few hospitals. Watson for Oncology has faced challenges due to its suboptimal concordance with human oncologists' recommendations, leading to subsequent distrust (Somashekhar *et al.*, 2018; Lee *et al.*, 2018b). It must be noted that OncoDoc, MATE, OncoCure, DESIREE, and WFO were designed o be used within cancer MTB meetings.
As a conclusion, one of the critical factors influencing the routine use of CDSSs is the acceptance by healthcare professionals. Factors such as effectiveness, ease of use, and user-friendly interfaces play a significant role in adoption. Another important factor is the seamless integration of CDSSs with EHRs, eliminating the need to re-enter patient data.

Overall, the successful implementation of CDSSs wether knowledge-based or non knowledge-based in clinical practice faces several technical challenges. Several studies (Patkar *et al.*, 2012; Séroussi *et al.*, 2013; MacLaughlin *et al.*, 2018) emphasized the integration of CDSSs with EHR systems, enabling efficient access to patient data, improving follow-up care, and enhancing guideline adherence. One crucial requirement is the semantic interoperability between CDSSs and electronic health record (EHR) systems, which is currently not adequately ensured (Sujansky, 2001). Differences in terminology and the handling of ambiguous or incomplete EHR data pose challenges to mapping patient data to CDSS functions (Gooch & Roudsari, 2011). Inconsistently stored EHR data and potential output generated by CDSSs have been identified as major barriers to CDSS implementation (Klarenbeek *et al.*, 2020a). Furthermore, the quality and reliability of clinical data in EHRs need to be interpreted and validated for reliable clinical decision-making, considering different sources and contexts of data entries (Jensen *et al.*, 2017; Ong *et al.*, 2017).

In the next two sections (section 2.2 and 2.3), a review of the approaches in guideline-based and

case-based reasoning will be presented. The tools used in Chapter 5 and the similarity measures employed in Chapter 6 will be described in detail.

2.2 Guideline-based reasoning

Clinical Practice Guidelines (CPGs) are textual documents developed by national agencies and healthcare societies to provide evidence-based recommendations for managing specific patient profiles (Bilici *et al.*, 2018). These recommendations are derived from clinical research findings and aim at representing the current state-of-the-art to support the practice of evidence-based medicine. The implementation of CPGs in clinical practice has been shown to improve decision-making quality by increasing adherence to recommended practices. In the case of breast cancer, adherence to CPG-recommended treatments has been found to correlate with improved patient survival rates (Voigt & Trautwein, 2023).
Despite these advantages, the actual adherence to CPGs in oncology remains suboptimal, as compliance with CPGs is a challenging topic because it depends on a variety of factors (Quaglini, 2008). Implementation barriers of CPG's are multifaceted (Voigt & Trautwein, 2023), including factors such as:

- Lack of knowledge and familiarity with CPG contents, as well as their limited applicability to specific clinical scenarios that are not consistently aligned with real-world decision-making practices.
- Textual presentation of CPG recommendations, which can be ambiguous and challenging to interpret.
- Dispersed nature of decision-relevant information within the CPG documents, making it difficult to extract the necessary elements for clinical decision-making for individual patient cases.

To address these challenges, guideline-based CDSS have been developed to assist healthcare professionals in their decision-making process and guide the clinical management of cancer patients by providing evidence-based recommendations (Klarenbeek *et al.*, 2020b). In the context of breast cancer, where medical complexity, time constraints, and the importance of informed decision-making in a shared decision-making environment are prevalent, automated CDSSs hold significant potential (Jiang *et al.*, 2019; Seroussi *et al.*, 2012a; Mazo *et al.*, 2020).

2.2.1 Formalization of guidelines

The formalization of CPGs can vary in depth depending on the editorial guidelines and the target audience. Venot (2013) distinguished different levels of formalization the goes from narrative CPGs to those more suitable for computer systems to process.

- Narrative textual CPGs: These CPGs are written in natural language and contain minimal or no structured elements. They tend to be lengthy and require careful interpretation.
- Semi-structured CPGs: These CPGs incorporate illustrations, tables, and decision trees to enhance clarity and reduce ambiguity. The text is written in a concise style using simple language.

- Structured CPGs: These CPGs are well-suited for computer systems. They go beyond generic knowledge representation formalisms and may include specific formalisms developed for CPGs like decision trees and knowledge graphs.

2.2.1.1 Computerized CPGs

Several models have been proposed to structure CPGs. The choice of representation language depends on the desired level of formalization and ease of execution (Quaglini & Ciccarese, 2006). We will briefly discuss historical formalisms used for CPGs. If you want to go further (De Clercq *et al.*, 2004) and (Peleg *et al.*, 2003; Peleg, 2013) provided a comprehensive review of these approaches.

- PROforma (Fox *et al.*, 1997), created at the Advanced Computation Laboratory of Cancer Research in the UK, is a modeling system that merges logic programming and object-oriented modeling, specifically grounded in the R2L Language. The project's goal was to investigate the capabilities of a minimalist set of modeling constructs. PROforma supports four key tasks: actions, compound plans, decisions, and inquiries, all of which share common attributes such as goals, control flow, preconditions, and postconditions.

- Arden Syntax: Developed by Jenders *et al.* (2003), Arden Syntax is a procedural language that allows the definition of conditional rules leading to actions. While it is easy for non-informatician physicians to create executable programs using Arden Syntax, it may lack flexibility for complex rules related to chronic disease management. Arden syntax have been adopted as the standard language by HL7 in 1999.

- GLIF: The Guideline Interchange Format, initiated by Ohno-Machado *et al.* (1998) in 1998, aimed to create a sharable model for CPGs using flowcharts. However, this model remained conceptual with no successful implementation.

- More recently Pournik *et al.* (2023) presented the CAREPATH methodology to develop computer interpretable, integrated clinical guidelines for managing multimorbid patients with mild cognitive impairment or mild dementia. The method involves three phases: conceptual modeling, interpretable modeling, and localization, emphasizing collaboration between clinical and technical teams.

As we can see, despite the early creation of common models, researchers have developed their own formalisms based on specific needs and goals. However, the trend shifted towards using standard technologies, particularly those derived from the semantic web, to improve the representation and sharing of CPGs (Peleg, 2013).

2.2.2 CDSSs and ontologies

The arrival of the Semantic Web and its standards in the early 2000s prompted a gradual evolution in decision support methods. Ontological reasoning, was seen as a valuable addition to traditional CDSS knowledge bases. Traditional CDSSs often lack the ability to adapt to knowledge and data due to cognitive rigidity. Ontological inference provides the capability to vary the level of abstraction.

2.2.2.1 Ontologies

In the context of semantic web, an ontology is a formal, explicit specification of a shared conceptualization. This means that an ontology is a way of representing knowledge about a particular domain.

The backbone of an ontology consists of a generalization/specialization hierarchy of concepts, i.e. a taxonomy. For example, in an ontology about human resources, the concept of 'Person' might be a superconcept of the concepts 'Manager' and 'Researcher'. This means that a 'Manager' and a 'Researcher' are both types of 'Persons'. In addition to the taxonomy, an ontology may also include other types of information, such as definitions of concepts, descriptions of relations between concepts, and constraints on the use of concepts (Guarino *et al.*, 2009).

According to Bodenreider (2008), ontologies can be used to facilitate knowledge sharing, data integration, and decision support in biomedical research. They have a number of benefits, including:

Improved knowledge sharing: Ontologies can help to ensure that researchers are using the same terms to refer to the same concepts, which can improve the communication and sharing of knowledge.

Enhanced data integration: Ontologies can be used to integrate data from different sources, which can help to identify relationships between data that would not be apparent otherwise.

Improved decision support: Ontologies can be used to represent clinical guidelines and decision rules, which can help healthcare professionals to make more informed decisions.

The use of ontologies in biomedical research is still in its early stages, but the potential benefits are significant.

2.2.2.2 Use of ontologies in CDSSs

Ontologies have been used for decision support in various domains, including patients with comorbidities, the COMET system, developed by Abidi (2011), integrates multiple clinical practice guidelines on heart failure and atrial fibrillation using ontologies. It creates a knowledge base and provides patient-centered recommendations. The system combines and fuses the guidelines to generate personalized recommendations for individual patients. Galopin (2015) created a methodology that was used in GO-DSS (Galopin *et al.*, 2015). It is a decision support system based on ontological reasoning for the flexible management of patients with multiple pathologies. It utilizes formalized clinical guidelines and an ontology-based patient profile graph to adapt knowledge to varying levels of patient description.

Regarding cancer, Abidi *et al.* (2007) introduced the Breast Cancer Follow-up Decision Support System (BCF-DSS). BCF-DSS combines decision rules from a clinical practice guideline with three ontologies (patient, breast cancer, and clinical practice guidelines) and a logical reasoning engine. The system uses backward chaining to offer personalized recommendations and justifications, however the system was not implemented or evaluated in practice.

Another project by Daniyal *et al.* (2009) uses ontologies to merge multiple clinical pathways for prostate cancer. The aim is to create a unified pathway as a foundation for a decision support system, but explicit ontological reasoning is not addressed in this case. An evaluation involving 10 physicians demonstrated the potential of such systems, emphasizing the importance of usability and interoperability in their design and implementation.

More recently, the DESIREE project aimed to create a guideline-based decision support system for breast cancer (Bouaud *et al.*, 2020b). This system is based on an ontology to build a model for breast cancer knowledge. We use this system as the guideline-based DSS for this project, more details about the ontology and the workflow of the systems will be presented in the next section (2.2.4).

2.2.3 Update of computerized knowledge bases

Since medical knowledge evolves constantly, CDSSs based on guidelines need to be updated regularly according to the most recent evidence. As we are concerned about updating a knowledge base in chapter 5. We did a scoping review to analyse the methods used for comparing narrative CPGs (Azarpira *et al.*, 2022).

A literature search was carried out using three distinct queries : (([computerized comparison] OR [manual comparison]) AND [clinical guidelines]), (computerized comparison of clinical guidelines), and (computerized evaluation of clinical guidelines). Searching PubMed and Google Scholar, we selected 11 relevant articles discussing automatic or semi-automatic methods for comparing CPGs. After analysing these articles, we considered three phases for comparing CPGs.

2.2.3.1 Concept Extraction :

The extraction of clinical concepts can be accomplished through different methods. One approach involves using rule-based techniques to search for exact word matches (Eftimov *et al.*, 2017). Alternatively, when the text lacks standard medical terminologies, neural networks with attention mechanisms can be employed (Tutubalina *et al.*, 2018; Gao *et al.*, 2021a). Extracted concepts are subsequently "normalized" using appropriate terminological dictionaries specific to the domain (Eftimov *et al.*, 2017; Tutubalina *et al.*, 2018). This process can be automated (Eftimov *et al.*, 2017) or rely on expert knowledge (Galopin *et al.*, 2014a).

2.2.3.2 Rule and Semantics Extraction :

In this phase, the goal is to extract recommendation statements and "if-then" rules to create a computer-interpretable version of narrative CPGs. Machine learning algorithms are commonly used for extracting recommendation statements with high accuracy scores (Hussain & Lee, 2019). While some studies have proposed using Natural Language Understanding (NLU) approaches to extract semantics and rules from CPGs (Schlegel *et al.*, 2019), the automatic extraction of "if-then" phrases from narrative text is still not entirely satisfactory. Currently, this phase requires the involvement of a human domain expert (Galopin *et al.*, 2015, 2014b).

2.2.3.3 Comparison of Guidelines at Different Data Abstraction Levels :

Comparing CPGs can be conducted at various levels :

- At the concept level, different methods based on the similarity between medical concepts have been evaluated against human expert assessments (Pedersen *et al.*, 2007). Mathematical techniques generally fall into two categories: (i) "ontological step-based" methods, which measure the minimal number of steps required to connect two concepts in an ontology, and (ii) "embedding matrices" that calculate similarity using metrics like Cosine similarity (Pedersen *et al.*, 2007; Zhu *et al.*, 2017) between the embeddings of the texts representing the guidelines.

- Other studies tried to do the comparison by evaluating the intersection of conceptual coverage in CPGs: Galopin *et al.* (2014a) have employed this method to compare three CPGs for managing arterial hypertension. And Mouazer *et al.* (2021) compared five CPGs for Potentially Inappropriate Medications (PIM) from different sources, utilizing visualization methods to illustrate the intersections of the conceptual coverage between CPGs.

- In another study, ontologies were utilized to map patient profiles and rules from two CPGs, enabling a comparison of inferred therapeutic recommendations. Inferred actions in such scenarios can sometimes be conflicting and may require resolution by a human expert (Galopin *et al.*, 2015, 2014b).

Overall updating CPGs is crucial for reliable guideline-based CDSSs. The process involves extracting and organizing concepts to create computer-interpretable CPGs. Concepts can be extracted using machine learning algorithms, rule-based NLP methods, or modern neural networks. Rule and recommendation extraction from CPGs still requires human domain expertise. CPG comparison can be done using ontologies, embedding matrices, or assessing conceptual coverage. Visualization methods also aid in intuitive comparisons and provide promising results. However, there is no fully satisfactory automatic approach for comparing CPGs.
In this work, chapter 5, we propose a new methodology to semi-automatically update of the GL-DSS of the DESIREE project according to the most recent evidence. The system is presented in the sections below.

2.2.4 Breast cancer knowledge model ontology

The Breast Cancer Knowledge Model (BCKM) ontology is at the core of the Guideline-based Clinical Decision Support System of the DESIREE project (Bouaud *et al.*, 2020b). It combines generic data model components with existing terminological resources to create an ontology to represent the knowledge and specifications related to breast cancer and its therapeutic management.
The ontology has been designed in accordance with the Entity-Attribute-Value (EAV) model, which is a generic model that is not specific to a particular domain. This model is flexible and allows for representing clinical data. The structure of the BCKM ontology is as follows:

Entities : In the BCKM, there are three main classes to describe a breast cancer patient state. These classes serve as descriptors for entities that play a role in describing a patient's case and are relevant in the decision-making process. While PatientEntity represent the patient herself, SideEntity is used to capture information related to the side affected, and LesionEntity describes any lesions present (refer to Fig 2.2 for a visual representation).
Apart from these entities directly linked to the patient case, there are other entities associated with the patient that help to contextualize their situation. These entities mainly encompass information about the patient's relatives, previous treatments they have undergone, and examinations carried out.

Attributes : The attribute classes list attributes of the different entities, e.g. the age of the patient, the BI-RADS value for a side, or the size of a lesion. In the BCKM ontology, each class of an attribute is declared to belong to an entity using the relation isAttributeOf. (e.g Age isAttributeOf PatientEntity)

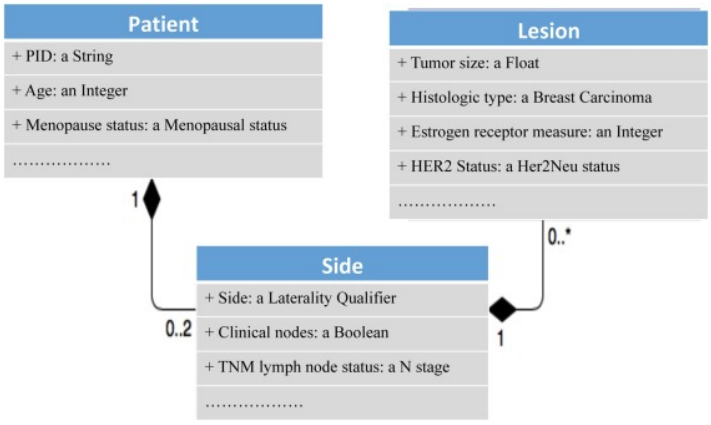


Figure 2.2: Excerpt from the UML class diagram representing the three main clinical entities (Patient, Side, and Lesion) used to describe a breast cancer clinical case, and their relationships (Bouaud *et al.*, 2020b)).

Values : These classes represent the different value types associated with attributes. These value types can include common primitive types like integers, floats, Booleans, dates, and strings. Additionally, there is a separate subclass called hierarchicalValue, which consists of discrete, potentially hierarchically organized by subsumption, value sets to describe values of attributes that are not primitive (e.g. the histologic type).
The specification of the value type for an attribute is achieved through the relation hasRange that establishes a link between the attribute class and the class representing the value type. For example, the Age attribute of a patient is linked via hasRange to the IntegerValue class, and the HistologicType attribute of a lesion is associated with the BreastCarcinoma class also via hasRange.

Figure 2.3 shows an overview of Ontology, a detailed presentation of all classes is provided in Appendix B.

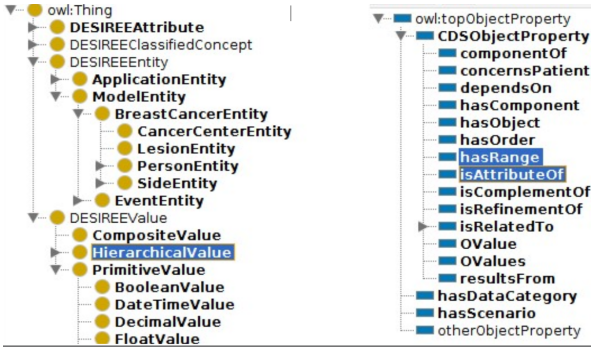


Figure 2.3: Overview of the BCKM ontology

Reasoning with the ontology : The BCKM ontology is used in the reasoning process of the GL-DSS (the process is explained in detail in the next subsection). Experts formalize rules using the

Natrual Rule Language (NRL) language and the Euler-Eye engine (Verborgh & De Roo, 2015) is used for rule-based reasoning and subsumption according to ontology.

The Euler-Eye engine is a rule-based reasoning engine that is used to process the rules in the BCKM ontology. The engine can reason about the ontology and infer new information from the existing information. This allows the GL-DSS to provide more personalized and evidence-based recommendations to breast cancer patients.

The NRL language is a language that is used to formalize the rules consistently with the EAV model that structures the BCKM ontology. The language is designed to be easy for experts to use and it allows the rules to be expressed in a natural way.

2.2.4.1 GL-DSS of the DESIREE project

The Guideline-Based Decision Support System is a system that provides personalized recommendations to patients based on clinical guidelines. Clinical guidelines are documents that describe the best practices for the management of a particular disease or condition. The GL-DSS supports a variety of clinical guidelines for breast cancer, including those from the US National Comprehensive Cancer Network (NCCN), the Spanish Onkologikoa group, and the French AP-HP hospital institution. Enabling the execution of the GL-DSS can be divided into three main steps:

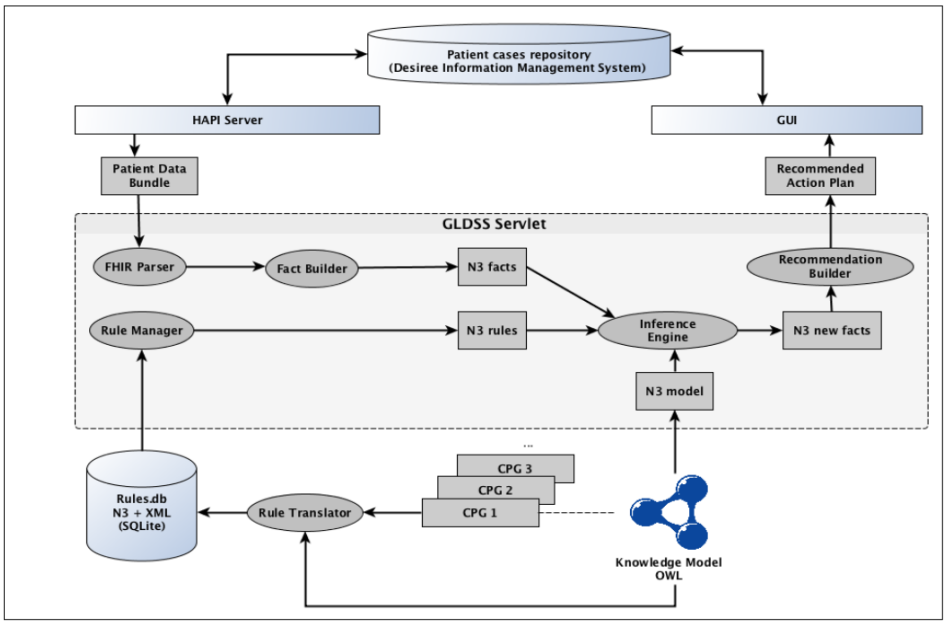


Figure 2.4: Workflow in the GL-DSS

1- Knowledge base creation : In this phase, in collaboration with the experts, the guidelines are formalized into an executable language by the computer (refer to the bottom part of figure 2.4 for a visual representation) :

1. The rules are first encoded in a formal language that is human-readable.
2. The NRL is then used to write the formal rules in a computer-readable language, using the concepts in the BCKM.

3. The NRL rule sets are then transformed into an XML representation, which is dependent on the BCKM ontology model.

4. Finally, a final transformation gives an triplet (N3) representation of the NRL rules, which is the target format that is interpretable by the Euler-EYE inference engine.

2- Patient data formalization This part focuses on transforming the patient's data represented according to the BCKM into triplets dependent on the BCKM ontology, to allow the Euler EYE engine to reason on the patient. The top left part of figure 2.4 provides a visual representation.

1. The initial patient data input is obtained via the DESIMS (Desiree Information Management System), which is the user interface used in the project. For development purposes, tools have been developed to also read patient data presented in a specific CSV format. The inputs are also dependent on the BCKM ontology, and the users fill in the values for the attributes of various entities (typically, the patient, side, and lesion entities).

2. Patient data is transferred into a standard FHIR bundle using FHIR messages (Patient, Observation, Condition, BodySite, and Specimen).

3. The FHIR bundle is transformed into a triplet representation (in N3 notation), which is consistent with the data model described in the BCKM.

3- GL-DSS execution: recommendation generation Once patient data is in N3 format, the rules (also transformed in N3 format) are matched to the patient's representation using the Euler-EYE engine. New facts are then created and finally, recommendations are provided. Recommendations are available in an XML format. Then, the DESIMS interface displays them to the user (top right part of figure 2.4). The figure below shows an example of the output of the GL-DSS, a file that shows the representation of the patient in a form of a knowledge graph, the facts inferred for this patient, and the recommendations derived from the APHP guideline.

In the Oncolog-IA project, we installed the GL-DSS in the AP-HP data warehouse environment and used it as the guideline-based CDSS in this project. In chapter 5, we describe how we automatically create patients in the BCKM format from the textual summaries used in MTBs and use the GL-DSS engine to provide recommendations for them.

2.3 Case-based reasoning

The medical reasoning process involves considering physiological conditions, patient complaints, symptoms, and other relevant factors when formulating treatment strategies (Lucas, 1993). Treatment reasoning necessitates cognitive activities such as information gathering, pattern recognition, and problem-solving in order to make informed decisions. Developing effective treatment plans can be a complex and error-prone task.

Case-Based Reasoning (CBR) (De Mantaras, 2001) is a machine learning research area based on the memory-centered cognitive model. It is an analogical reasoning method that involves reasoning from past cases or experiences to solve problems or interpret anomalous situations. CBR integrates problem-solving, understanding, and learning by utilizing memory processes (De Mantaras, 2001). It is characterized by adapting previous solutions to address new demands, using old cases to explain new solutions, and reasoning from past events to interpret novel situations. CBR

is akin to similarity-based reasoning, as it assumes that similar problems have similar solutions (Armengol *et al.*, 2004).

According to Koton, "A physician's problem-solving performance improves with experience. The performance of most medical expert systems does not" (Koton, 1989). The implication is that, instead of using rules to decide, clinicians rely on their acquired knowledge from books and experiences, which is very similar to the functioning of Case-Based Reasoning (Schmidt *et al.*, 2001). CBR systems offer a significant advantage through the automatic formation of a facility-adapted knowledge base (Schmidt *et al.*, 1999). This adaptability is crucial in medical decision-making processes. Additionally, the dynamic nature of medical knowledge, the existence of multiple solutions, and the complexity of modeling make CBR applicable and relevant in the medical domain (Holt *et al.*, 2005). Consequently, CBR has been widely employed in the development of intelligent computer-aided decision support systems within the medical field over the past few decades (Ahmed *et al.*, 2010).

Several models have been proposed to elucidate CBR functioning. These models include Hunt's model, Allen's model, and others (Leake, 1996). Among these models, the R4 model by Aamodt & Plaza (1994) is the most widely adopted and provides a high-level and comprehensive framework (Finnie & Sun, 2003). The process depicted in this model can be represented by a schematic cycle consisting of the four R's: Retrieve the most similar cases, Reuse the information and knowledge from retrieved cases to solve the problem, Revise the proposed solution, and Retain the part of this solution likely to be useful in the future, as illustrated in Figure 2.5.

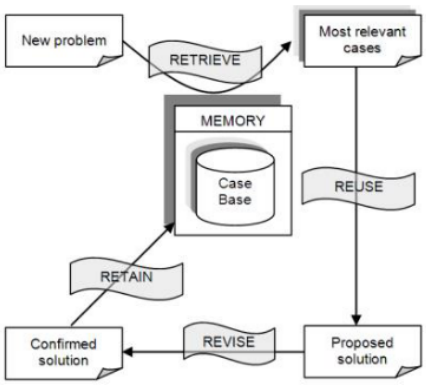


Figure 2.5: The R4 cycle (Aamodt & Plaza, 1994)

Among the four R's, Retrieval is considered the most important part, as it sets the foundation of CBR systems (De Mantaras *et al.*, 2005). Retrieval includes the process of finding the most similar cases to the current case. The most commonly used techniques include nearest neighbor retrieval, inductive approaches, and knowledge-guided approaches (Pal & Shiu, 2004; Simoudis & Miller, 1990). In Chapter 6, we propose a method for similarity calculation for patients with breast cancer. In the next subsections, we will review the state-of-the-art similarity calculation methods.

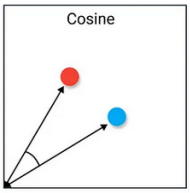
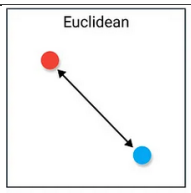
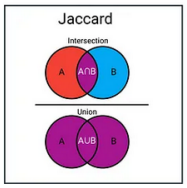
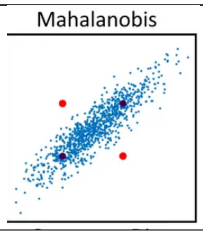
Choudhury & Ara (2016) proposed a survey on CBR in medicine that dive deeper into the other parts of CBR systems and their application in healthcare for diagnosis and classification.

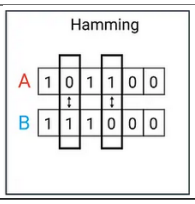
2.3.1 Similarity measures

Similarity measures between patients can be used to identify subgroups of patients with similar clinical profiles, predict individual patient outcomes, and inform personalized treatment decisions (Parimbelli *et al.*, 2018). By identifying subgroups of patients with similar clinical profiles, clinicians can develop more targeted and effective treatment strategies. Evaluating patient similarity has been explored as a means to facilitate precision medicine and has been recognized as a crucial problem in many data mining algorithms and real-world information processing systems.

We did a literature review on similarity measures used for patient similarity and noticed there are two main ways of evaluating the similarity between two patients. Some works use Deep Learning models, others classical Machine Learning methods such as linear regression, K-nearest neighbors... Both methods rely on various similarity measures (also called metrics) to calculate the similarity of pairs of patients. Table 2.1 reports on the metrics used in most studies. We will then describe the different methods.

Table 2.1: Summary of papers by each similarity measure

Measure	Definition	Image	Articles
Cosine similarity	Cosine similarity is a measure of similarity between two vectors that computes the Cosine of the angle between them. A Cosine similarity value close to 1 indicates that the vectors are very similar, while a value close to 0 indicates that they are very different.		(Pai *et al.*, 2019; Lee *et al.*, 2018a; Wang *et al.*, 2019; Chen *et al.*, 2022)
Euclidean distance	Euclidean distance is a measure of distance between two points in an Euclidean space. It is defined as the length of the straight line segment connecting the two points.		(Pai *et al.*, 2019; Lee *et al.*, 2018a; Wang *et al.*, 2019; Pai & Bader, 2018)
Jaccard similarity	Jaccard similarity is a measure of similarity between two sets that computes the ratio of the number of elements common to the two sets to the total number of elements in the two sets.		(Pai *et al.*, 2019; Pai & Bader, 2018; Meystre *et al.*, 2019; Huang *et al.*, 2021)
Mahalanobis distance	Mahalanobis distance is a measure of distance between two points in an Euclidean space that takes into account the variance and standard deviation of the data.		(Panahiazar *et al.*, 2015; Oei *et al.*, 2021)

Hamming distance	Hamming distance is a measure of distance between two strings that counts the number of locations where the two strings differ.		(Lee *et al.*, 2018a)

Lee *et al.* (2015) utilized a patient similarity method based on Cosine similarity for mortality prediction. Panahiazar *et al.* (2015) developed two distinct methodologies for suggesting medications to patients with heart failure. These methodologies involved clustering techniques such as hierarchical clustering and K-means clustering using the Mahalanobis distance, as well as classification techniques. Their study demonstrated that classification was more effective than clustering in improving medical diagnosis, personalized treatment planning, and prediction accuracy. Li *et al.* (2015) introduced an unsupervised clustering methodology based on topological examination to detect subgroups of type 2 diabetes patients. This methodology aimed to identify distinct subgroups within the diabetes population without prior supervision.

A new software **NetDx** Pai *et al.* (2019) has been developed for the construction of interpretable patient classifiers by integrating multi-omics and structured EHR data using patient similarity networks (Pai & Bader, 2018). The authors have applied numerous similarity measures (including Jaccard, Cosine, and Euclidean) to construct networks of similar patients and integrate them into a common network.

Overall, various metrics or similarity measures are commonly employed in the retrieval phase. Many studies focus on directly learning these metrics to calculate the similarity between patients (Choudhury & Ara, 2016) and subsequently create clusters using methods such as k-nearest neighbors (kNN) or weighted kNN (Lamy *et al.*, 2019). More recently (Gérardin *et al.*, 2022a) worked on the development of a cohort for patient similarity from French clinical notes, based on automatic concepts extraction, they test their approach on 6 phenotypes and showed promising results using Earth mover's distance (Kusner *et al.*, 2015). Other recent works explored the use of the clinical notes text directly to detect similarity between patients, using state-of-the-art language models (van Aken *et al.*, 2022).

Metric learning, in general, involves the task of learning a similarity measure that reflects the desired notion of similarity or dissimilarity between data points. Traditional metric learning approaches typically rely on defining a parametric form of the metric and learning its parameters based on labeled or pairwise similarity information. However, a recent advance in the field is the emergence of deep metric learning methods.

2.3.2 Deep metric learning

Deep metric learning specifically leverages deep neural networks to learn effective representations for metric learning (Suo *et al.*, 2018). Unlike traditional methods that rely on handcrafted features or metrics, deep metric learning automatically learns high-level representations from raw data. This enables the capture of complex and non-linear relationships in the data, which may be challenging for traditional methods.

In the context of patient similarity assessment, deep metric learning has been applied to learn patient representations from electronic health records (EHRs) (Suo *et al.*, 2018). Authors used Convolutional neural networks (CNNs) to capture important information from EHRs, and the learned representations are fed into loss functions such as triplet loss or softmax cross-entropy loss. Exper-

imental results have shown that deep metric learning improves the representation of longitudinal EHR sequences and outperforms traditional distance metric learning methods.

In our work, chapter 6, we use a deep metric learning method to learn embeddings on our patient's data. As we perform the triplet loss function as an objective function, we explain triplet loss in the next subsection.

2.3.2.1 Triplet loss

Triplet loss (Hoffer & Ailon, 2018) is a loss function used in deep metric learning to learn effective embeddings or representations of data points. It operates on triplets of data points: an anchor point, a positive point, and a negative point. The objective of triplet loss is to minimize the distance between the anchor and positive points in the embedding space while maximizing the distance between the anchor and negative points, by predefined margins. By optimizing the triplet loss, the neural network learns to map similar data points closer together in the embedding space and push dissimilar data points further apart. This enables the network to capture meaningful similarities and dissimilarities between data points, facilitating tasks such as clustering, retrieval, and classification.

Triplet loss has been widely applied in various domains, including face recognition, person re-identification, and healthcare. In healthcare, triplet loss (Schroff *et al.*, 2015) can be used to learn patient embeddings based on their medical records and clinical features, enabling the identification of similar patients for personalized treatment recommendations and decision-making.

Now that we described the various approaches used in clinical decision support for oncology and had an overview of the tools and methods we will be using in this work. Let's focus on how we get the inputs for these CDSSs, as we mentioned in the introduction, we use textual unstructured clinical notes as inputs and we have to process and structure them to use them in CDSSs. The next section focuses on state-of-the-art natural language processing methods.

2.4 Natural language processing in healthcare

Natural language processing has become an important domain in the medical informatics field. NLP enables the analysis and interpretation of textual data, which is heavily used in healthcare for traceability and communication, thus facilitating information retrieval, knowledge extraction, and decision support. In this section, we explore various aspects of NLP, including computer representations of text, named entity recognition, and describe some NLP tools that we used in our work.

2.4.1 Computer representations of text

Textual data in healthcare presents unique challenges that require appropriate computer representations for effective exploitation. Several approaches have been developed to represent text computationally, enabling downstream NLP tasks.
In NLP, texts are divided into smaller units called tokens. Tokenization can be performed at different levels, such as word, character, or subword, depending on the specific requirements of the task. The choice of tokenization level impacts the system's generalizability and performance.

After tokenizing a sentence into individual words, each word is associated with a set of features. In early NLP methods, hand-engineered features such as word cases, punctuation patterns,

part-of-speech (POS) labels, and various linguistic properties, were specifically designed to capture relevant information tailored to specific NLP tasks. To delve deeper into such features, one can refer to the detailed review published by Nadeau & Sekine (2007).

Some systems also relied on terminologies like UMLS (Bodenreider, 2004), acting as dictionaries with diverse expression representations based on different characteristics. Textual entity identification involved exact matching or distance calculations between text fragments and terminological entries at word or character levels.

2.4.1.1 Word embeddings

Early NLP methods relied on word embeddings are sets of real-valued features associated with words (Collobert & Weston, 2008). They can be learned from scratch or computed from morphological features (Klein *et al.*, 2003; Bojanowski *et al.*, 2017; Akbik *et al.*, 2018). Word embeddings capture the implicit semantics of words and have become the standard for analyzing text with machine learning (Collobert & Weston, 2008).

Pretrained Representations Pretrained representations have gained popularity in NLP. Static word embeddings have been trained on tasks such as language modeling (Mikolov *et al.*, 2013; Turian *et al.*, 2010; Collobert *et al.*, 2011). Language modeling involves capturing the distributional properties of words, guided by the idea that "*a word is characterized by the company it keeps*" (Firth, 1957; Harris, 1954). Word2Vec (Mikolov *et al.*, 2013) and GLOVE (Pennington *et al.*, 2014a) are examples of static word embeddings. FastText (Bojanowski *et al.*, 2017) is another variant that represents words using character n-grams.

However, static word embeddings do not consider the context of a word when used in a new sentence, which limits their usefulness for certain cases such as homonyms or referent words like pronouns.

To overcome this limitation, contextualized word embeddings have been introduced. ELMO (Peters *et al.*, 2018) improved upon static embeddings by pretraining a deep recurrent language model and using hidden representations as features for downstream tasks. BERT (Devlin *et al.*, 2019) is another popular model that uses masked language modeling. There have been various modifications and variants proposed, both in terms of model architecture and pretraining corpus domain (Clark *et al.*, 2020; Kong *et al.*, 2020; Liu *et al.*, 2019b; Yang *et al.*, 2019; Martin *et al.*, 2020; Beltagy *et al.*, 2020a; Lee *et al.*, 2020). **HuggingFace** (Wolf *et al.*, 2020) has played a significant role in popularizing these models by simplifying their implementation and facilitating sharing.

For a comprehensive review of the research field, Qiu *et al.* (2020) provides a detailed analysis of contextualized word embeddings.

2.4.1.2 Large Language Models

NLP is now driven by large language models. These models, such as GPT and BART, are trained on massive amounts of text data and can perform a wide range of NLP tasks, including language understanding, generation, and translation. Large language models serve as a common backbone for various NLP applications, eliminating the need for task-specific architectures and achieving state-of-the-art performance. (Lewis *et al.*, 2020; Radford *et al.*, 2018, 2020; Raffel *et al.*, 2019; Brown *et al.*, 2020a). However, concerns related to their size, biases, and ethical implications have also been raised (Bender *et al.*, 2021).

2.4.2 Named Entity Recognition

Named Entity Recognition (NER) is a natural language processing technique that involves identifying and classifying named entities, such as names of diseases, measurements, dates, and other specific entities, within a given text. NER plays a crucial role in information extraction and helps in understanding the relationships and context between different entities in a text. It has undergone several developments since its emergence in the early 1990s. Initially, NER focused on rule-based systems that relied on handcrafted rules, lexical functions, and gazetteer lists (Rau, 1990; Brin, 1999; Collins & Singer, 1999; Riloff & Jones, 1999; Lin, 1998; Alfonseca & Manandhar, 2002; Etzioni *et al.*, 2005). These systems used annotated data and employed heuristics and rules for generalization. Early works explored different techniques to enhance entity detection.

A significant revolution in NER came with the introduction of sequence labeling systems, which treated NER as a word classification problem. Tags were assigned to individual words to indicate their position within an entity, and various tag schemes were introduced to represent entity boundaries and types (Huang *et al.*, 2015; Klein *et al.*, 2003; Lample *et al.*, 2016). Supervised methods like Decision Trees, Support Vector Machines, and Conditional Random Fields (CRF) were commonly used.

As NER progressed, researchers turned their attention to handling more complex scenarios such as nested and overlapping entities. Nested NER, which deals with overlapping entities, gained attention with the GENIA corpus (Kim *et al.*, 2003). Different methods were proposed to address various aspects of nested entities, including constituency parsing graph extraction (Finkel & Manning, 2009), CRF hyper-graph modeling (Lu & Roth, 2015), mention edges and transitions (Muis & Lu, 2017), and multi-label prediction (Katiyar & Cardie, 2018). Exhaustive NER methods, which enumerate all possible spans in the input sequence, were also developed (Xu *et al.*, 2017; Wang *et al.*, 2020; Zheng *et al.*, 2019; Luan *et al.*, 2019).

Advancements in NER include alternative formulations such as sequence-to-set approaches and machine reading comprehension tasks. These approaches leverage pre-trained language models and transfer learning to predict overlapping entities without explicit type annotation. Other techniques incorporate deep language models with markup tags or combine models to extract flat, nested, and overlapping entities (Tan *et al.*, 2021; Li *et al.*, 2020; Mengge *et al.*, 2020; De Cao *et al.*, 2020; Yan *et al.*, 2021).

2.4.2.1 Recent works on clinical NER and relation extraction

A recent review titled *Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review* by Fraile Navarro *et al.* (2023) examined the current literature on clinical NLP systems that perform multi-entity NER and Relation Extraction (RE). The review aimed to provide an update on the state-of-the-art performance, explore clinical task descriptions, assess real-world implementation, and highlight areas for improvement.

The authors found that recent developments in NER and RE heavily utilized pre-trained BERT-derived models, specifically tailored to the biomedical domain, such as BLUEBERT (Peng *et al.*, 2019) and Pub-MedBERT (Gu *et al.*, 2020) and those specifically developed with clinical text, like ClinicalBERT (Alsentzer *et al.*, 2019), and EHRBERT (Li *et al.*, 2019). However, the availability of these models for implementation was limited due to proprietary datasets and a lack of publicly available resources. The review highlighted the need for more validation and deployment of NLP systems in real-world clinical settings.

Although newer systems such as GPT-3 (Brown *et al.*, 2020b) have been popular with chatGPT,

they seem to fail when evaluated in medical use cases (Rousseau *et al.*, 2020). Developers willing to implement or develop NLP clinical systems can reflect on the availability of these tools to avoid costly training and, thanks to novel transfer learning techniques and easy-to-use implementations (Wolf *et al.*, 2019).

The analysis of the included studies revealed that a significant number of them did not describe specific clinical tasks or problems to be solved. Only a small portion addressed a single clinical or information task, such as adverse event detection and pharmacovigilance, increasing clinician and patient understanding, decision support, drug efficacy, coding and automating EHR tasks, quality improvement, public health, and genotype and phenotype research. Additionally, the review identified that only a few studies provided evidence of deployment in real-world settings. The majority of the included studies (72%) employ Machine Learning and deep learning methods for extracting entities and relationships from text, showing promise in capturing complex patterns and improving performance. However, the review showed that rule-based methods are still used, offering interpretable extraction rules and effectiveness in scenarios with consistent patterns and limited labeled data. Overall, hybrid approaches (24% of the included studies) combining rule-based and ML methods provide robust and accurate extraction systems.

2.4.3 NLP tools for healthcare

Many NLP tools have been developed for structured information extraction from, annotation, and curation of biomedical or clinical texts. These tools facilitate various tasks such as literature review, disease-centered relation extraction, and biomedical concept recognition. Among these tools we can distinguish :

Annotation tools are tools that are generally graphical interfaces used to manually annotate documents, these tools facilitate the annotation process to annotate data for machine learning algorithms and also to evaluate them.

Information extraction tools : These tools serve as systems to extract relevant entities from the text automatically, these entities can be, clinical concepts like UMLS codes, drugs, or ontology concepts ..etc.

2.4.3.1 Clinical text annotation

Annotator (Shah *et al.*, 2009), **Argo** (Rak *et al.*, 2012) and **GATE Teamware** (Cunningham *et al.*, 2002), provide annotation and extraction capabilities for biomedical text. Annotator and Gate offer functionalities to retrieve annotations using ontologies, while Argo incorporates text-mining techniques for biocuration workflows.
BioNotate (Cano *et al.*, 2009) and **BeCAS** (Nunes *et al.*, 2013) support relation extraction and concept recognition in biomedical text. BioNotate facilitates the annotation of disease-centered relations.
BioQRator (Kwon *et al.*, 2013) and Tagtog (Cejuela *et al.*, 2014), offer platforms for biomedical literature curation. They provide interactive interfaces and efficient environments for annotating texts Tagtol is designed to detecet gene mentions in PLOS.

In this work we use **BRAT** (Stenetorp *et al.*, 2012a) : It is a widely used, annotation tool and online environment for collaborative text annotation. It provides an efficient interface for annotating and curating text, supporting distant collaborative annotation tasks and streamlining the

annotation process. We use it since it is the annotation tool installed withing the environment of the AP-HP datawarehouse.

2.4.3.2 Clinical information extraction

SemRep (Rindflesch & Fiszman, 2003) extracts semantic predications from biomedical free text by identifying subject-relation-object triples. It captures meaningful information for advanced analyses.
Both Apache **cTAKES** (Savova *et al.*, 2010) and **MetaMap** (Aronson & Lang, 2010) extract UMLS concepts. cTAKES specializes in analyzing clinical text in healthcare settings for information extraction and provide customizable information extraction pipelines. MetaMap maps biomedical text to UMLS concepts, facilitating the identification and analysis of medical concepts.
EventMine (Miwa *et al.*, 2012) focuses on extracting semantically enriched events from biomedical literature, enabling researchers to identify and analyze specific events of interest.
CLAMP (Soysal *et al.*, 2018) is a Clinical Language Annotation, Modeling, and Processing tool designed for healthcare and clinical text. It offers a toolkit for efficiently building customized clinical natural language processing pipelines.

In this work (4.2.2), we used MetaMap as an automatic annotator to extract UMLS concepts from BCPSs translated from French to English.

2.4.3.3 NLP for cancer

Numerous studies have focused on extracting structured information from cancer reports. Various methods have been developed to automatically extract features from clinical reports, such as radiology reports (Bitterman *et al.*, 2021; Miwa *et al.*, 2014) and pathology reports (Burger *et al.*, 2016). Some research has concentrated on specific attributes within reports, treating the task as a classification or term extraction problem for items like BIRADS scores, histological grade, or lesion site (Alawad *et al.*, 2018; He *et al.*, 2017; Moore *et al.*, 2017; Qiu *et al.*, 2018). Specialized systems have also been designed for features like locations (Datta *et al.*, 2020), and a comprehensive survey of these systems is presented by Datta *et al.* (2019).

Other studies have aimed to achieve more comprehensive and global extraction, simultaneously detecting multiple entity types. The earliest approach by Taira *et al.* (2001) introduced a frame-based representation for annotating abnormal findings, anatomy, and medical procedures in radiology reports. Lacson *et al.* (2015) employed a rule-based system and terminologies to extract abnormal findings and BIRADS scores.

The DeepPhe system (Savova *et al.*, 2017) was developed as an integrated software using cTakes (Savova *et al.*, 2010). DeepPhe extracts cancer "summaries" from clinical reports, encompassing pathology, radiology, and observations.

2.4.3.4 Tools for French language

Despite recent improvements in natural language models, machine understanding of language, especially clinical documents in French, is still far from being solved. English has a wide range of processing tools and terminology resources that surpass those available for other languages. Not all English approaches can be directly applied to French due to language-specific differences. Additionally, while there is considerable research on general domain texts in French, the biomedical domain lags behind (Nêvêol *et al.*, 2018). For example, despite being the 5th most represented

language in the 2019 version of the UMLS terminology, French only has synonyms for 3.5% of its concepts. Therefore, developing methods for clinical NLP in French is a crucial aspect of this work.

Although most of the described tools and methods primarily focus on English text, there have been some advancements in the field of clinical NLP for the French language. Machine learning methods have been employed to extract clinical information from texts, and there are also available semantic annotators.

Semantic Annotators Two main systems are used as semantic annotators for the French language:

1. **SIFR** (Semantic Indexing of French Biomedical Data Resources) (Tchechmedjiev *et al.*, 2018) is an openly available web service that facilitates the recognition and contextualization of concepts from 30 medical terminologies and ontologies. The annotator service processes textual descriptions, tags them with relevant biomedical ontology concepts, including UMLS, expands the annotations using the knowledge embedded in ontologies, and contextualizes the annotations before returning them to the users in various formats.

2. **ECMT** (Extracteur de Concepts Multi-Terminologique) is a web service inspired by the CISMef algorithm for information retrieval with the Doc'CISMeF search engine and F-MTI, which is a multi-terminology automatic indexer (Pereira *et al.*, 2008). ECMT is designed for the French language and offers two query modules: a default module based on a bag-of-words algorithm (Pereira *et al.*, 2008), and an expanded module based on textual indexing using Oracle Text Indexing. The ECMT annotator works with seven terminologies and supports semantic expansion features (Sakji *et al.*, 2010).

In this work, section 4.2.2, we used ECMT as an automatic annotator to extract clinical concepts from the BCPSs.

Information extraction from clinical notes Several studies have addressed information extraction from clinical notes in French. Grouin *et al.* (2011) developed a system to compute the CHA2DS2-VASc score, assessing stroke risk in patients with non-valvular fibrillation. Digan *et al.* (2019) investigated the impact of text duplications on clinical narratives, emphasizing the importance of identifying and annotating duplicated zones. Jouffroy *et al.* (2021) developed a hybrid deep learning system for medication-related information extraction. Lerner *et al.* (2020) proposed a terminologies augmented recurrent neural network model for clinical named entity recognition.
Wajsbürt (2021) focused on the extraction and normalization of simple and structured entities in medical documents, contributing to improving entity extraction in French clinical texts, and the creation of an NLP library: **NLstruct** for nested NER (Wajsbürt *et al.*, 2021; Gérardin *et al.*, 2022b).
More recently Aouina *et al.* (2023) developed an ontology-based semantic annotation system for French psychiatric clinical documents, enabling the identification of important medical events and the construction of patient profiles.

A work has been also recently done on extracting relevant data for breast cancer from clinical notes. Schiappa *et al.* (2022) developed an automatic tool, called RUBY, for structuring clinical information from French medical records of patients with breast cancer. RUBY combines state-of-the-art NER models with keyword extraction and postprocessing rules. The tool achieved high precision rates ranging from 81.8% to 98.1% in extracting specific information from different types of clinical reports. The information extracted in this work is close to the ones we want to extract in

chapter 3, but since the system is not publically available, we used their paper as a comparison to our results.

EDS NLP As we work on clinical notes stored in the AP-HP data warehouse, we used **EDS-NLP** (Dura *et al.*, 2022b). The EDS-NLP is a powerful tool developed by the AP-HP (Assistance Publique - Hôpitaux de Paris) data science team specifically designed for extracting valuable information from clinical notes written in French. It comprises several components including text normalization and entity extraction, qualifiers such as negation detection and family context detection, as well as miscellaneous components for date extraction and measurements. Additionally, it incorporates specific components for named entity recognition.

The NER pipeline presented in chapter 4 relies mainly on EDS-NLP. The system was used to implement custom pipeline using regular expressions. More details about the components can be found at `https://aphp.github.io/edsnlp/latest/`.

2.4.4 Clinical text classification

As we are facing a text classification problem in chapter 4 to classify breast cancer patient summaries as complex or non-complex. In this section we describe different methods used to do clinical document classification. We can resume the pipeline for classifying a clinical documents into 3 steps starting with a preprocessing step, the feature engineering step to extract relevant information, and finally a classification step using machine learning-based or rule-based techniques.

We will describe the types of clinical reports used in various published studies, then we will go into the details of preprocessing and sampling techniques. We will also explore the feature engineering techniques employer and also the machine learning algorithms that are mentioned in the literature.

2.4.4.1 Types of clinical reports

Clinical text classification techniques have been widely used in analyzing various types of free-text clinical reports. These reports include pathology reports, radiology reports, autopsy reports, death certificates, and biomedical documents in general. Among these, pathology reports were most commonly utilized (Mujtaba *et al.*, 2019), followed by radiology reports, and autopsy reports. Pathology reports were primarily employed to detect breast cancer and other related cancers using text classification techniques. For example, Rani *et al.* (2015) utilized pathology reports to identify cancer stages through text classification. Similarly, Kasthurirathne *et al.* (2017) investigated the detection of cancer from pathology reports using non-dictionary-based and dictionary-based text classification approaches.

Radiology reports were also extensively utilized in clinical text classification.Zuccon *et al.* (2013) employed radiology reports to identify limb fractures through text classification techniques. Shin *et al.* (2017) used radiology reports related to brain computed tomography to identify pediatric traumatic brain injury. Bates *et al.* (2015) employed radiology reports to detect HIV using automated text classification.

In addition, researchers have classified influenza-related clinical reports to detect influenza-like illnesses using supervised machine learning (Pineda *et al.*, 2015; Ye *et al.*, 2014). Death certificates and autopsy reports have also been utilized to determine the cause of death (Butt *et al.*, 2013; Danso *et al.*, 2013; Mujtaba *et al.*, 2018)

Recent studies have focused on combining various clinical reports related to the same disease to develop classification models. For example, Kavuluru *et al.* (2015) combined pathology and radiology reports to automatically assign ICD-9 codes to electronic medical reports. Kocbek *et al.* (2016) combined pathology reports, radiology reports, and patients' admission-related metadata to predict admission rates for specific diseases. Combining data from different sources or combining features of different reports has been found to improve the reliability and accuracy of predictions in these studies.

2.4.4.2 Preprocessing

In the preprocessing stage of clinical text classification, various techniques were applied to the narrative clinical reports. These techniques included spell correction, tokenization, stemming, lemmatization, and normalization. For example, numerous studies (Danso *et al.*, 2013; Lauren *et al.*, 2017; Sarker & Gonzalez, 2015; Buchan *et al.*, 2017; Wang *et al.*, 2017; Clark *et al.*, 2017; Martinez *et al.*, 2015; Masino *et al.*, 2016) implemented these preprocessing tasks.

Stop word removal, removal of punctuation and white spaces, and case conversion were commonly used as basic preprocessing tasks. Word tokenization was also widely employed. Some studies investigated the impact of stop words and found that their presence improved classification accuracy. Stemming, when combined with basic preprocessing tasks and word tokenization, was found to enhance classification performance in several studies (Adeva *et al.*, 2014; Jo, 2013; Koopman *et al.*, 2015; Zuccon *et al.*, 2015; Sarker & Gonzalez, 2015). Additionally, researchers such as applied stemming and lemmatization techniques for clinical text normalization, reporting their effectiveness in improving classification performance. Text normalization techniques using regular expressions were used to convert numbers or dates to common units, such as number and date, thus addressing the issue of dimensionality. Studies (Buchan *et al.*, 2017; Wang *et al.*, 2017) demonstrated the effectiveness of these text normalization techniques.

2.4.4.3 Feature engineering

Feature engineering in text classification involves feature extraction, feature value representation, and feature selection. Various studies have investigated these steps in the context of clinical text classification. In the feature extraction step, two general approaches have been explored: expert-driven and fully automated feature extraction.
Fully automated feature extraction techniques involve extracting content-based features (such as Bag of Words (BoW), n-gram, and Word2Vec (Goldberg & Levy, 2014)), concept-based features (using medical terminologies like SNOMED-CT (Donnelly, 2006; Stearns *et al.*, 2001; Spackman *et al.*, 1997)), structural features (utilizing the structure of clinical documents), linguistic features (including parts of speech), and graph-based features or graph of word (GoW) features.
Expert-driven feature extraction relies on domain experts to manually extract relevant features from clinical reports (Sedghi *et al.*, 2016). Experts rank the extracted features based on their discriminative power and store them in lexicons for classification.

The choice of feature representation techniques also plays a role in text classification. Binary representation (BR), term frequency (TF), term frequency with inverse document frequency (TFiDF), and normalized TFiDF are commonly used techniques (Debole & Sebastiani, 2004). Each technique assigns a numeric value to each feature. BR represents features as binary values (0 or 1), TF represents features based on their frequency in a document, TFiDF considers the frequency of a feature in the document and its rarity across the dataset, and N-TFiDF combines term frequency,

document frequency, and document length to ensure equal importance for features in long and short documents.

Feature selection techniques are employed to identify the most relevant subset of features. Information Gain (IG), Chi-square (chi), Pearson Correlation (PC), Local Semi-Supervised Feature Selection (LSFS), Expert-driven (ED) ranking, Mutual Information (MI), Gini-Index (GI), Distinguishing Feature Selector (DFS), Principal Component Analysis (PCA), Multiple Discriminant Analysis (MDA), and Bi-Normal Separation Score (BNSS) are some of the techniques used in clinical text classification (Yang & Pedersen, 1997; Benesty *et al.*, 2009; King *et al.*, 2010; Guyon & Elisseeff, 2003; Loh, 2011; Uysal & Gunal, 2012; Forman, 2003)

2.4.4.4 Machine learning and rule-based classification

In the field of clinical text classification, both machine learning and rule-based classification approaches have been applied. Machine learning algorithms utilize historical data to learn patterns and make predictions, while rule-based classifiers rely on manually crafted rules to perform classification tasks. Machine learning algorithms, such as support vector machine (SVM), naive Bayes (NB), decision trees (DT), random forest (RF), k-nearest neighbors (kNN), and artificial neural networks (ANN), have been extensively used in clinical text classification (Rani *et al.*, 2015; Kasthurirathne *et al.*, 2016; Zuccon *et al.*, 2013; Shin *et al.*, 2017; Bates *et al.*, 2015; Pineda *et al.*, 2015; Mujtaba *et al.*, 2018). SVM has shown high performance in several studies, often outperforming other algorithms (Butt *et al.*, 2013; Kasthurirathne *et al.*, 2016), but the other machine learning algorithms, such as NB, RF, DT, and kNN, have also been successfully employed in various clinical text classification tasks.

Rule-based classifiers have been used as an alternative approach in clinical text classification. These classifiers rely on manually defined rules or expert knowledge to classify documents (Alghoson, 2014; Deng *et al.*, 2015; Koopman *et al.*, 2015; Kalter *et al.*, 2016). Rule-based classifiers are flexible and easy to understand, and misclassification errors can be corrected more easily compared to machine learning approaches. However, they heavily depend on the expertise of rule designers and may lack scalability.

In our work, we employed a combination of machine learning and rule-based techniques for feature extraction on the task of text classification.

2.5 Conclusion

In this thesis, we aim to address some of the main problems identified in and presented in the literature review above.

- In Chapter 3, a pipeline is developed to extract relevant data in a structured form directly from patients' clinical notes. This automated data extraction process is proposed as a solution to avoid the need for manually setting the inputs of the CDSS. We implement several NLP tools and methods described above to create a robust pipeline for structured data extraction.

- Chapter 4 focuses on implementing a pipeline to distinguish between simple and complex patients. Simple cases are those covered by existing guidelines, while complex cases require a more individualized approach. For complex cases, a case-based decision support system is proposed to provide tailored recommendations. Using this method, we can capture the

patients covered by the CPGs, thus dealing with the problem of uncommon clinical profiles that causes nonadherence to guideline-based CDSS.

- In Chapter 5 we automatically create patient profiles within the guideline-based CDSS for simple cases identified by the methodology described in chapter 4. Leveraging the extracted structured data from chapter 3, this system offers evidence-based recommendations and treatment options based on established guidelines. In this chapter we propose an efficient method to semi-automatically update the knowledge base of a guideline-based CDSS, dealing with the problem of regularly updating knowledge bases.
- Chapter 6 addresses the complex cases using a case-based CDSS. This system calculates similarities between patients and provides personalized decision support based on similar cases.

CHAPTER 3

Data extraction from textual breast cancer patient summaries

Before learning case complexity and providing decision support according to the complexity, we need to have structured data that can be used as input to learning the complexity of each clinical case, but also structured data that can be mapped to the guideline-based decision support system of DESIREE and can be also used to implement similarity measures for the case-based reasoning. . The effective extraction of structured data from clinical notes is a critical aspect of advancing healthcare research and clinical decision support systems. In this chapter, we investigate the efficiency of a rule-based method for data extraction from breast cancer patient summaries (BCPSs) and explore the associated challenges and limitations.

The rule-based pipeline demonstrated very good performance overall, achieving perfect results on attributes for which precise rules were tailored, emphasizing the importance of customization for improved precision, recall, and F1-scores. Intriguingly, this evaluation also led to the identification of attributes that were not considered in the annotation process, underscoring the dynamic nature of clinical data and the need for pipeline adaptability. The pipeline showed also very good performance regarding relation extraction and contextual information extraction.

Despite facing challenges like delayed data access and inconsistencies in BCPS data within the data warehouse, we addressed these limitations through rigorous data cleaning and processing. As a result, the rule-based method proved to be effective in handling BCPSs. The research in this chapter underscores the significance of data quality management in data-driven healthcare applications and demonstrates the positive impact of customized rules.

In concluding this chapter, we highlight the potential of a hybrid approach that integrates rule-based techniques with deep learning methods to address the dynamic nature of clinical data and tackle complex concepts effectively. Embracing this hybrid methodology and investing in annotated data can lead the pipeline to evolve into a robust tool capable of precise and comprehensive context extraction, significantly enhancing its applicability and value in clinical settings.

3.1 Introduction

In the field of healthcare and medical research, electronic medical records (EMRs) play a crucial role as a valuable source of patient information. EMRs contain essential data such as medical history, diagnoses, laboratory results, imagery reports, discharge summaries, orders, treatments, and more. These comprehensive records have the potential to significantly improve the quality of care delivered to patients and support evidence-based decision-making processes. However, the lack of standardization in EMRs poses challenges in terms of efficient reuse and effective querying of their content. However, most of the information content of a patient record is provided as text (Raghavan *et al.*, 2014). Extracting accurate and relevant information from narrative notes often requires labor-intensive medical record reviews, impeding the seamless integration of computer-based processing tasks like decision support.

Over the past few decades, artificial intelligence (AI) techniques, specifically natural language processing (NLP), have been employed in oncology research with varying degrees of success. Coden *et al.* (2009) demonstrated effective results in extracting various information (histology, location, primary tumor, etc.). However, when parsing pathology reports, data extraction proved unsatisfactory due to the inherent nature of clinical notes. These notes lack standardized language, exhibit ambiguous abbreviations, diverge from common reporting guidelines between clinicians and organizations, and may contain complex temporal relationships (Wieneke *et al.*, 2015; Forsyth *et al.*, 2018).
In France, the promotion of medical record utilization has been a key objective over the past decade, exemplified by initiatives such as the Health Data Hub program, which aim to accelerate the digital transition of the healthcare system (Plantier *et al.*, 2017). However, advanced tools dedicated to oncology remain scarce.

With the objective of extracting structured data from clinical notes, we (Redjdal *et al.*, 2022a) tried to use automatic semantic annotators such as those described in chapter 4.2.2.1. Even though these systems were useful to produce outputs used as features to machine learning algorithms for complexity classification (see chapter 4), we noticed that using semantic annotators does not allow to extraction of all specific data related to breast cancer. For instance, numerical values like the hormonal receptors, the tumor size, or the number of positive lymph nodes are rarely detected by annotators. Other important features like the **TNM** staging or the histological type of tumors are also often missed by annotators.

In this chapter, we focus on the extraction of structured data from breast cancer patient summaries. The objective is to develop a structured representation that supports complexity learning (chapter 4), case-based decision support (chapter 5), and a mapping with the BCKM ontology for guideline-based decision support (chapter 6).

To achieve this, based on the entity attribute model value (EAV) used in the BCKM ontology, we created a pipeline, integrating rule-based techniques to annotate phrases of patient summaries that refer to structured data elements. Rule-based methods serve as a pre-annotation tool, while we plan that machine learning methods will be employed for concepts where rule-based methods are inefficient.

The extraction process is guided by the task and will prioritize capturing all information utilized by clinicians in making decisions for patients. In addition to providing structured data for the decision support system, such algorithms can be used within the AP-HP data warehouse, as a valuable tool for extracting structured data from breast cancer patient summaries. This automated approach not only saves time but also ensures the consistency and reliability of the extracted data, facilitating robust analysis and further insights into cancer patients' care.

3.2 Material and methods

As we use the guideline-based decision support system of the DESIREE project (described in section 2.2.4.1), we adopt the structured data model provided by the BCKM ontology (described in section 2.2.4) as the main data model in the structured data extraction task.

In this section, we describe the pipeline we created, starting from unstructured clinical documents, using natural language processing techniques to obtain structured data presented in the form of a patient data graph.

3.2.1 Breast cancer patient summaries and structured data model

3.2.1.1 Breast cancer patient summaries

As described in the introduction, this project has been approved by the institutional review board at AP-HP (CSE 200094). We obtained access to the data of pseudonymized patients diagnosed with breast cancer between 2018 and 2022. We had access to a sample of 3,500 patients, each patient containing one or more breast cancer patient summaries (BCPSs), available as textual unstructured documents (the number of BCPSs varies from 1 to 4 or more, with an average of 3 BCPSs per patient). BCPSs provide a portrait of patients with all the relevant information that MTB clinicians need to know to make the best patient-specific therapeutic decision.
A typical BCPS (see figure 3.1) contains all the information needed to decide the best care plan given a breast cancer patient. Information is organized following the order below:

- Personal information (available as unidentified in the data warehouse)
- Biometric data
- Reason for presentation
- Personal history (medical and surgical history, followed treatments and allergies.)
- Family history
- History of the disease
- Clinical examination
- Radiology results (Mammography, Echography, MRI ..)
- Biopsy results
- TNM classification
- Response to neoadjuvant treatment (if any)
- Pathology results (if prior surgery was done)
- Treatment proposal

However, all BCPSs do not contain all the information above, and many of them do not follow the structure above. This unstructured format makes information extraction complicated. There is a lack of standardized language, the use of many abbreviations, acronyms, and specialized terms. A variety of terms may be used, that may not correspond to a general domain, depending on the health professional specialty of the BCPS's author.

Données Médicales : Motif de présentation Validation simple de RCP
Histoire de la maladie
Patiente âgée de [age], adressée par le [MEDECIN EXTERIEUR]pour un Cancer canalaire infiltrant du sein droit.
Antécédents:
Chirurgicaux: aucune
Gynécologiques: Ménopausée depuis l'âge de 53ans, pas de THM.
Anamnèse: Découverte d'une lésion sur bilan de dépistage. Pas de douleur, pas d'écoulement.
Mammographie/ Echographie de dépistage [36 jours avant RCP] [VILLE]:
Sein droit: ACR 4c avec masse uni-focale de 10mm du QIE à 7h30 à 1cm du mamelon.
Sein gauche: ACR1
Pas d'adénopathie axillaire.
Microbiopsie [28 jours avant RCP] au Raincy sein droit: Carcinome canalaire infiltrant NST, grade 2 EE 3+2+2, RO 100%, RP 100%; HER 2 score 1 négatif, Ki67% 15%.
Examen clinique : SG 100C sein droit : nodule centimétrique juxta aréolaire externe sein gauche : pas de lésion palpable aires ganglionnaires axillaires sus et sous claviculaires libres
Explications données sur le diagnostic et les modalités de traitement.
Indication de tumorectomie du sein droit après repérage et technique du ganglion sentinelle droit
BLOC opératoire le 08.03.22 en J0.
Brassière.
Décision de la RCP :
Proposition de décision
Proposition de décision
On confirme

Figure 3.1: Example of a de-identified breast cancer patient summary.

3.2.1.2 Structured data representation

As mentioned above, we use the EAV model. It is a data model that efficiently represents entities by leveraging a sparse matrix-like structure. It is designed to handle situations where there are numerous attributes that could potentially describe entities, but only a limited number of attributes are applicable to each specific entity. The EAV model is also known as the Object-Attribute-Value model, Vertical Database model, and Open Schema (Nadkarni *et al.*, 1999).

In contrast to information models specifically designed for the biomedical domain, such as OpenEHR, OMOP, and FHIR, which provide predefined objects tailored for hospital information systems or electronic health records, the EAV model is a generic model that can be applied to various domains and is considered to have the flexibility necessary to handle biomedical data, as noted by Khan *et al.* (2014); Löper *et al.* (2013); Nadkarni *et al.* (1999).
From a logical perspective, data models, whether they are relational or object-oriented, can be mapped or transformed into the EAV model.
Therefore, following the organization of the BCKM ontology made by Bouaud *et al.* (2020b), we organized the target database based on the components of the BCKM model. (Please refer to section 2.2.4 for a detailed description of the ontology and the model).

In line with the BCKM model, we collaborated closely with oncology experts who possess extensive knowledge and experience in the field. Through their expertise, we were able to discern the most crucial attributes relevant to each of the main entities. We also determined the potential values that could be assigned to these attributes. In the subsequent description, we provide a comprehensive account of each characteristic:

Patient characteristics

- Age
- Menopausal status (premenopausal vs. postmenopausal)
- BRCA (if clinically indicated) and other genetic mutations (Yes vs No)
- Comorbidities (especially important for geriatric patients and for identifying complex patients)
- Body mass index
- Response assessment to neoadjuvant therapy (disease progression, stable disease, partial response, complete response)
- **Oncotype Dx** (high risk, intermediate risk, low risk)
- Bilateral breast cancer (yes vs. no)

Side characteristics

- Multifocal breast cancer (Yes vs No)
- Birads score (0, 1, 2, 3, 4, 4a, 4b, 4c, 5, 6)
- Widespread microcalcifications (yes vs no)
- BraSize and cup
- TNM classification

Tumor characteristics

- Histology (ductal, lobular, others)

- Tumor size
- **Grade** (Grade1,Grade2,Grade3)
- Ki67 prognosis factor (numerical value in %)
- Breast cancer subtype (Hormone receptor positive/ HER2-negative, Hormone receptor positive/ HER2 positive, Hormone receptor negative/HER2 positive, Triple negative)
- SISH or FISH for HER2 in case of HER2 ++ (positive or negative)

Based on the attributes described above, we proposed an annotation scheme to model the structured data extraction algorithm. In the next section, we detail the annotation scheme and the resulting dataset. The relevant characteristics to extract were the result of discussions with the MTB physicians. The annotation scheme itself was the result of many iterations between annotations and scheme revision.

3.2.2 Annotation scheme

We first detail the annotation scheme. We focus on entities cited in the previous section. For the annotation, we use BRAT annotation tool (Stenetorp *et al.*, 2012b). Figure 3.2 shows the expert-annotated version of figure 3.1.

3.2.2.1 Entity annotation

In BRAT we don't have an explicit patient entity mention as the text itself refers to the patient. We use the entity mentions to annotate the side and the lesion entities. Each entity is annotated when there is a mention of the lesion for the lesion entity and when there is a mention of a breast side laterality for the side entity *(e.g Mammographie à* ***droite****<SideEntity>:* ***lesion****<LesionEntity> de 3cm)*.

In addition to the three main entities of the model (patient, side, lesion), we added the annotation of treatment and diagnosis procedures. This information can be useful when making a decision, especially information about past treatment procedures. Treatment and diagnosis entities are expressed in BRAT as entity mentions, and some of them have attributes representing the potential values of the entity (e.g. macro and micro biopsy for the biopsy entity). The annotation scheme for these is presented in the table 3.1:

Table 3.1: Diagnosis and treatment procedures in the annotation scheme

Entities	**Values**
Treatment Procedures	
Surgery	Mastectomy, Breast reconstruction, Breast plastic surgery, Conservative surgery, Axillary dissection, Sentinel lymph node Biopsy, Breast re-excision, Nodes re-excision, Annexectomy
Radiotherapy	Chest Wall, Boost, Sus claviculaire, Tumor bed irradiation.
Chemotherapy	Neo-adjuvant or adjuvant.
Endocrine Therapy	Neo-adjuvant or adjuvant.

Anti-HER2 therapy	Single or dual blocked.
Diagnotic procedures	
Biopsy	Mircobiopsy, Macrobiopsy.
Cytoponction	Positive, Negative.
Ultrasound	Text.
Mammography	Text.
Pet scan	Text.
Clinical examination	Text.

3.2.2.2 Attribute and value annotation

In the BRAT tool, only 4 types of mentions can be used, ***the entity mention, the attribute mention, the event mention, and the relation mention***. Considering this model, we had to adapt the annotation scheme to fit into it. Therefore patient characteristics are annotated as entity mentions and their values are expressed as attribute mentions. In addition to that, BRAT does not allow the user to put a textual or integer value for an attribute, all possible values of an attribute must be put into the annotation configuration file. This is why in the annotation scheme, for attributes that have integer values like the tumor size or text values like comorbidities we expressed as a value the text annotated. For attributes with hierarchical and boolean values, the value of each attribute is expressed using an attribute mention.

For example in figure 3.3, the attribute menopausal status corresponds to an entity mention in BRAT, and has the value Postmenopausal that corresponds to an attribute mention in BRAT. The attribute BraSizeCup corresponds to an entity mention in BRAT and has the value 105B which is the text annotated. The table 3.2 shows all the attributes and their values in the annotation scheme.

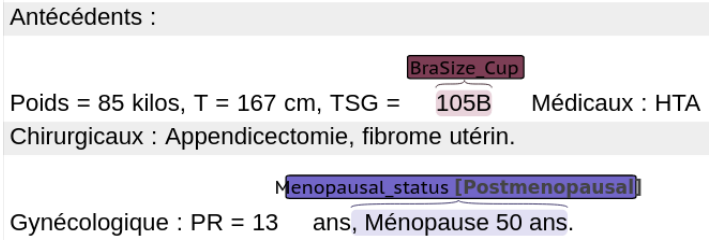


Figure 3.3: Example of attribute and value annotation

Table 3.2: Attributes of the main entities and their values in the annotation scheme

Attributes	**Values**
Patient entity	
Age	Integer
Menopausal status	Premonopausal, Postmenopausal, Perimenopausal
BRCA and other genetic mutations	True, False
Comorbidities	Text
Antecedent of another cancer	Breast cancer, Other cancer, Any cancer
Body mass index	Integer
Response assessment to neoadjuvant therapy	Disease progression, Stable disease, Partial response, Complete response

Oncotype Dx	Text (RS score)
Pregnancy Status	Pregnant, Desired pregnancy
Bra size and cup	Text
Treatments	Text

Side entity	
Side laterality	Left, Right, Bilateral
BIRADS classification	0, 1, 2, 3, 4, 4a, 4b, 4c, 5, 6
Confirmed positive nodes	True, False
Clinical positive nodes	True, False
Widespread microcalcifications	True, False
TNM	Text
N status	Text
Clear surgical margins	True, False
Cavity shave margin	Upper, Upper outer, Upper inner, Lower, Lower inner, Lower outer, Inner, Outer, Lateral
Cancer stage	Text
Node size	Integer
Max distance between tumors	Integer

Lesion entity	
Tumor	Text
Histology	Invasive Breast Carcinoma, Breast Sarcoma, Lobular Breast Carcinoma, Invasive Ductal Breast Carcinoma, Invasive Ductal and Lobular breast carcinoma, Invasive Lobular Breast Carcinoma, InSitu Breast Carcinoma, DCIS Breast Carcinoma, Lobular InSitu Breast Carcinoma, Non-cancer, Other
Associated InSitu carcinoma	True, False
Presence of emboly	True, False
Tumor site	upper inner quadrant, axillary region, areolar region, upper outer quadrant, lower outer quadrant, lower inner quadrant, central, union inner quadrant, union outer quadrant, union upper quadrant, union lower quadrant, under the mammary fold, mastectomy scar
Tumor size	Integer
Tumor grade inv	Grade1, Grade2, Grade3
Tumor grade insitu	Low Grade, Intermediate Grade, High Grade
Estrogen receptor	Integer
Progesterone receptor	Integer
HER2 status	Text
Ki67	Text
FISH	Text

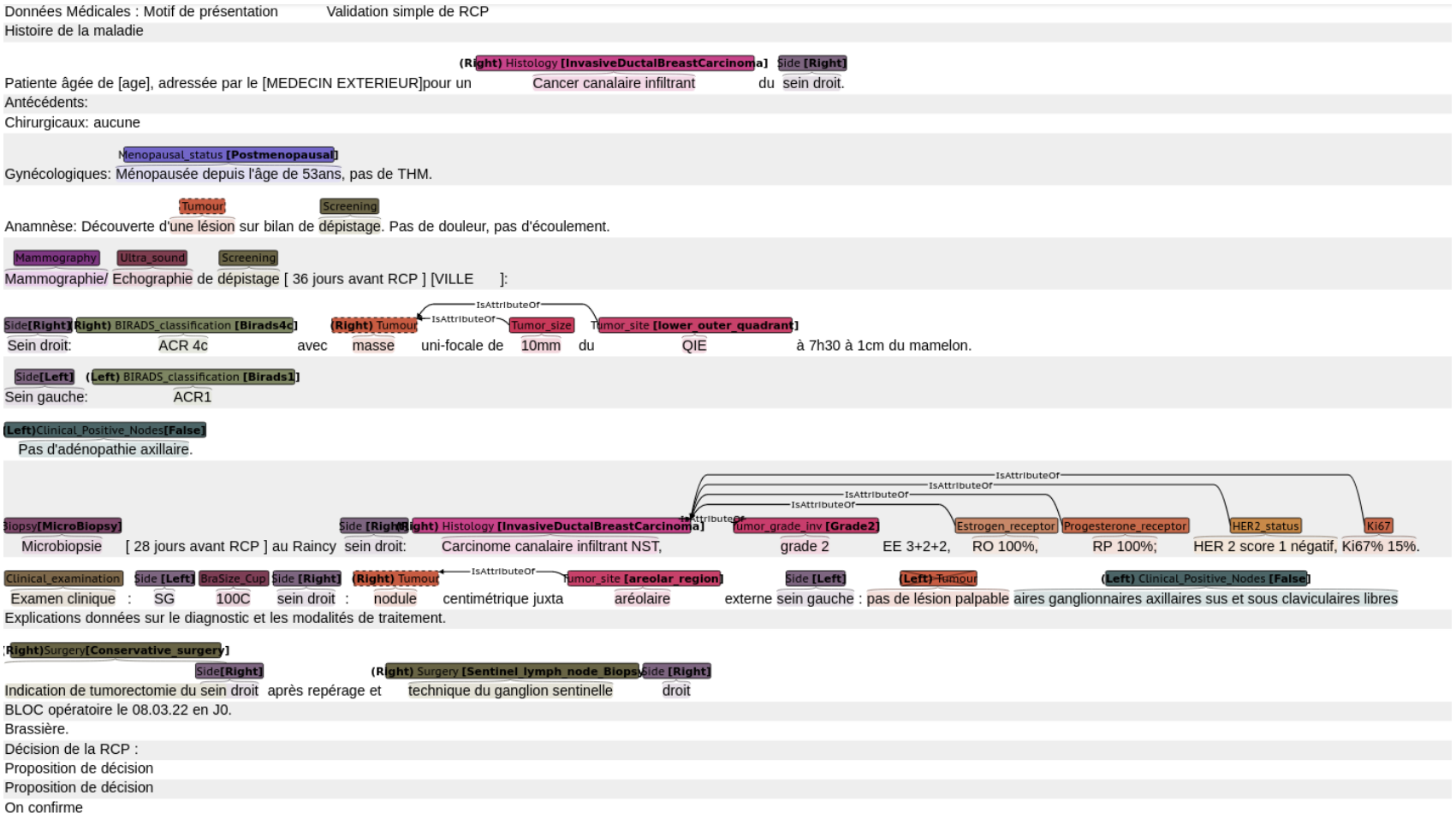


Figure 3.2: Annotated version of Figure 3.1.

3.2.2.3 Contextual information annotation

Extracting contextual information in the clinical domain is of paramount importance as it provides crucial insights that enhance the understanding and interpretation of medical data extracted from clinical notes. In BRAT we express the contextual information as attribute mentions, that can be applied to various entities to detect: the negated entities, the hypothetical entities, the family-related entities, the patient preferences, and the antecedent entities.

For example, in Figure 3.4:

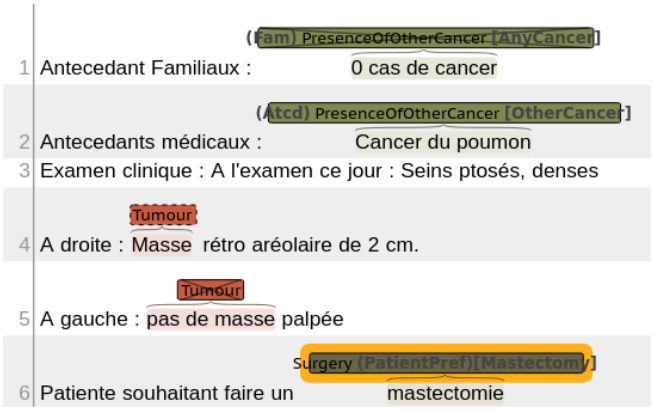


Figure 3.4: Contextual information annotation

- 1st line, the entity antecedent of another cancer (referred to as PresenceOfOtherCancer in the figure) is negated and affected to the family because the text reads ***Family history****: 0 cases of cancer*.
- Line 2, the entity antecedent of another cancer is tagged as antecedent because the text reads ***Medical history****: Lung cancer* .
- Line 4, the entity tumor is hypothetical because it is a clinical examination.
- Line 6, the entity mastectomy is tagged as patient preference because the text says ***Patient wishing*** *to have a mastectomy*.

3.2.2.4 Relations annotation

Since BRAT was not originally designed to annotate long multi-line relations, we tried to use as less relation mentions as possible. Following the entity, attribute, and value model. there are 3 types of relations in this work:

- **Has side :** This relation expresses the relation between a lesion entity and a side entity. To avoid a multi-line relation we express this relation using the attribute mention (left, right, or bilateral) that can be applied to any lesion entity or histologic type
- **Is attribute of :** This relation is used to link the attributes of an entity with the entity itself. In this work, we need to express the relation between the side attributes and the side entity. And also the relation between the lesion attributes and the lesion entity.
  As the side attributes can be related to the left or/and the right side, we keep the same methodology that we used to express the has side relation, as we mention the relation between the attribute and the side using the attribute mentions left, right, or bilateral.

  For the tumor attributes, we use a relation name isAttributeOf to express the link between a tumor attribute and its tumor entity. Even if we still have some problems with the multi-line relations, we notice that the tumor attributes are generally expressed just after the mention of the tumor entity, so while all the relations are expressed, the documents can still be readable using this relation as we can see in figure 3.5.

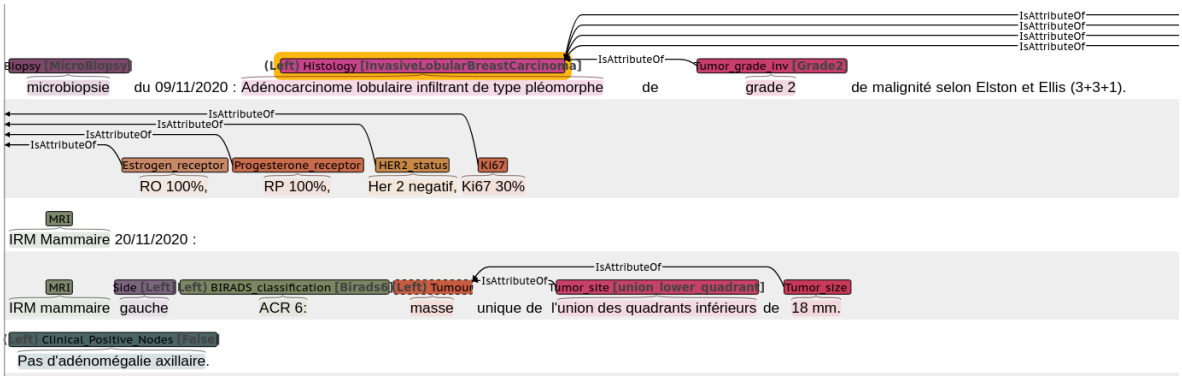


Figure 3.5: Relation annotation

3.2.3 Breast cancer named entity recognition

Given the time-consuming nature of manual annotation for machine learning, a rule-based pre-annotation model can be immensely helpful. The proposed approach begins with pre-annotating clinical notes using rule-based NLP techniques, utilizing domain-specific rules and regular expressions to identify and tag the specific entities mentioned in the annotation scheme within the text.

The utilization of rule-based NLP brings forth numerous advantages. Firstly, it provides a rapid and cost-effective means of developing annotation guidelines, as domain experts can actively contribute to the creation of rules based on their expertise. This reduces the dependence on extensive manual annotation, which tends to be a time-consuming and resource-intensive process. Moreover, rule-based methods promote transparency and interoperability, as the rules can be thoroughly reviewed, refined, and adjusted to enhance performance. This adaptability facilitates an iterative refinement of the NER process, ensuring the generation of high-quality annotations. In this section, we describe the pipeline we developed to implement the named entity recognition.

3.2.3.1 Text processing

We leveraged the capabilities of EDS-NLP (Dura *et al.*, 2022b), a powerful tool developed by the AP-HP data science team (see section 2.4.3.4.3), to extract the relevant information in French clinical notes. Since none of the EDS-NLP tools was adapted to provide satisfactory results, we designed a specialized pipeline within the EDS-NLP library for extracting breast cancer-related information from BCPSs.

Normalization We used the EDS-NLP normalization pipeline. It follows a non-destructive normalization approach, ensuring that the original input text remains unaltered. The normalizer operates according to four dimensions, including converting the text to lowercase, removing accents while maintaining character length, normalizing apostrophes and quotation marks, detecting and marking spaces and new lines.
To optimize the normalization process, we developed an end-of-line detection algorithm. In this algorithm, we used the end-lines model provided by EDS-NLP instance, which is an unsupervised algorithm based on the work of Zweigenbaum *et al.* (2016). In addition to that we implemented a rule-based approach to make more accurate endlines detections (e.g. if a line starts with an uppercase and the line before ends with a point, THEN this is an end of the line).

3.2.3.2 Rule-based named entity recognition

After the normalization process, we worked on extracting the attributes for each entity and their mentions. This process was a mix of using pipelines already implemented within EDS-NLP, regular expressions and using **Clarity NLP** (Georgia_Research_Institute, 2018)) for attributes where results were better than regular expressions and EDS-NLP components (attributes are the TNM classification and the Tumor size). Table 3.3 below shows the methods used to extract every mention in the text. The attributes and entities extracted correspond to the ones described in table 3.2:

Table 3.3: Methods used to extract attributes and their values from the text

Entities or attributes	Methods
Patient entity	
Age	Available as structured data
Menopausal status	Regular expression + postprocessing rules
BRCA and other genetic mutations	Regular expression + postprocessing rules
Comorbidities	Regular expression + Semantic annotators (ECMT) + ICD10 codes function from EDS-NLP
Antecedent of another cancer	Regular expression + Semantic annotator (ECMT) + ICD10 codes extraction function from EDS-NLP
Body mass index	Available as structured data
Response assessment to neoadjuvant therapy	Regular expression
Oncotype Dx	Regular expression
Bra size and cup	Regular expression
Pregnancy Status	Regular expression
Treatments	Drugs extraction function from EDS-NLP
Side entity	
Side laterality	Regular expression
BIRADS classification	Regular expression
Confirmed positive nodes	Regular expression + postprocessing rules
Clinical positive nodes	Regular expression + postprocessing rules
Widespread microcalcifications	Regular expression
TNM	Clarity NLP's TNM component
N status	Regular expression
Clear surgical margins	Regulars expression + postprocessing rules
Cavity shave margin	Regulars expression
Cancer stage	Regulars expression
Node size	Clarity NLP's size component + postprocessing rules
Max distance between tumors	Clarity NLP's size component + postprocessing rules

Lesion entity	
Tumor	Regular expressions
Histology	Regular expressions + semantic annotator ECMT
Associated InSitu Carcinoma	Regular expression + postprocessing rules
Presence of Emboly	Regular expression + postprocessing rules
Tumor site	Regular expressions
Tumor size	Clarity NLP's size component + postprocessing rules
Tumor grade inv	Regular expression
Tumor grade insitu	Regular expression
Estrogen receptor	Regular expression
Progesterone receptor	Regular expression
HER2 status	Regular expression
Ki67	Regular expression
FISH	Regular expression

3.2.3.3 Contextual information identification

As mentioned in section 3.2.2.3, the aim is to enhance the accuracy and reliability of information extraction from clinical notes. To do this, we use EDS-NLP components, and enrich them with a regular expression to identify the context:

Negation identification: The eds.negation pipeline employs a simple rule-based algorithm, inspired by the NegEx algorithm developed by Chapman *et al.* (2001), to detect negated spans within breast cancer patient summaries.
The pipeline achieved a notable Negation F1-score of 71% for CAS/ESSAI (Grabar *et al.*, 2018) and 88% for NegParHyp (Dalloux *et al.*, 2017), indicating its effectiveness in detecting negated information.

Family history identification: Similar to eds.negation, the eds.family pipeline utilizes a rule-based algorithm to identify spans or tokens within the text that refer to family members or family history rather than to the patient herself.

Hypothesis identification: The eds.hypothesis pipeline employs a rule-based algorithm to identify speculative spans within the text. These speculative spans denote information that is not certain but rather represents hypotheses or potential assumptions.
The pipeline achieved a Hypothesis F1-score of 49% for CAS/ESSAI (Grabar *et al.*, 2018) and 52% for NegParHyp (Dalloux *et al.*, 2017), indicating its ability to identify speculative information within breast cancer patient summaries.

Patient preference identification: EDS-NLP does not include a pipeline to identify patient preferences. However, this information is really important within the domain (e.g. if a patient wants a mastectomy while the guidelines say lumpectomy, doctors have to take into consideration the patient's preference and consider doing the mastectomy). That is why we use reg-

ular expressions to detect if a patient refuses another treatment then the one recommended by the guidelines.

3.2.4 Structured data extraction

Once we had a named entity extraction (NER) algorithm to annotate entities within the text. We developed a structured data extraction pipeline to capture important information about a patient. The pipeline involves several steps to ensure accurate extraction and organization of data.

First, we identify relevant sections within the text that contain information about the patient. These sections may include details about medical history, symptoms, treatments, and other pertinent factors.

Next, we utilize the NER algorithm described in 3.2.3 to annotate entities within the text. This helps us identify specific pieces of information such as medical conditions, medications, and procedures.

Once the entities are annotated, we extract relations between the annotated entities. After extracting the relations, we filter the data according to the patient's pathway. By doing so, we create a final structured version of the breast cancer patient summary that is consistent and easily usable in a CDSS.

For a visual representation of the structured data extraction pipeline, please refer to figure 3.6

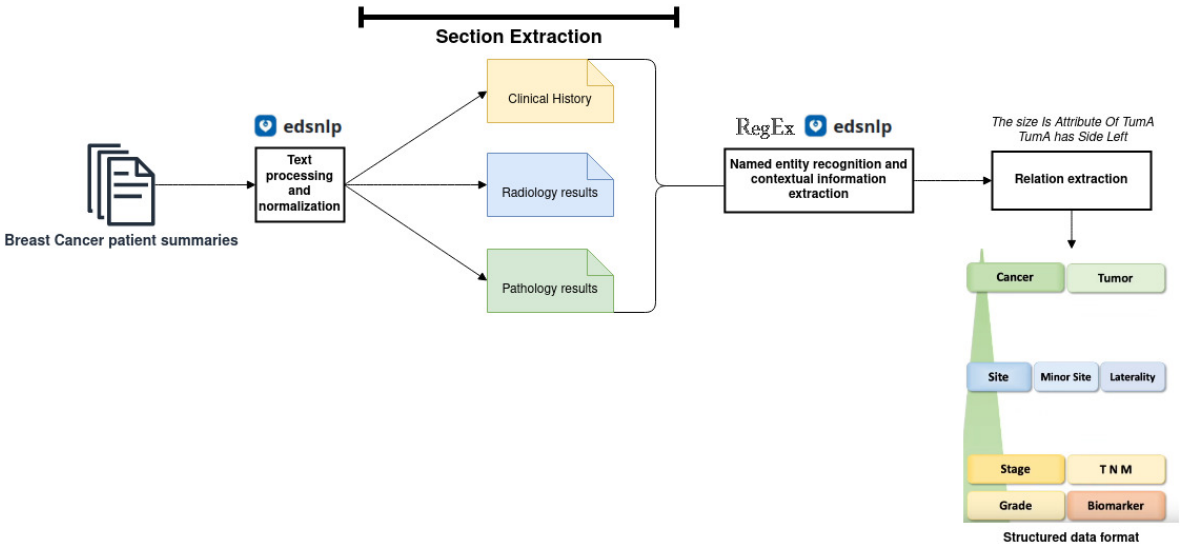


Figure 3.6: Structured data extraction pipeline

3.2.4.1 Section identification

In order to extract structured data from clinical notes for breast cancer patients, a first and crucial step is the identification and extraction of relevant sections within the text. To accomplish this, a rule-based approach is employed. The process begins with a step to pre-possess the text as explained in 3.2.3.1.1. This ensures that the subsequent analysis can be performed accurately.

To do the section extraction, we created a set of predefined section markers or keywords that are indicative of specific sections within the clinical notes. These section markers are carefully se-

lected based on domain knowledge and an understanding of the typical structure of breast cancer patient summaries as mentioned in 3.2.1.1. For instance, section markers such as "Personal Information," "Biometric Data," "Radiology Results," and "Treatment Proposal" are used to identify the corresponding sections within the BCPS.

Once the section markers are defined, the algorithm scans the preprocessed clinical notes for occurrences of these markers. It employs rule-based matching using regular expressions to identify the beginning of each section based on the proximity of the markers and the surrounding context. For instance, if the section marker "MRI Results" is found, the algorithm looks for the nearest section marker or a distinctive pattern that indicates the start of a new section (e.g. "Biopsy Results"). The new section indicates the end of the last section.

In cases where section markers alone are insufficient to accurately identify the boundaries of sections, additional linguistic patterns and contextual information can be utilized. For example, the algorithm may take into account the presence of specific keywords or phrases that commonly appear at the start or end of certain sections (e.g. clinicians always use "MTB of [a date]" to introduce the conclusion of previous MTBs.

Upon identifying the boundaries of sections, the algorithm extracts the corresponding text, including any relevant subsections or subheadings. Then the NER algorithm presented in 3.2.3 to extract the entities for each section with their contextual information. Finally, to build a patient timeline, we extract for each section, date mentions using eds.dates, a component of EDS-NLP. Using this method, sections are then organized into a structured format, facilitating the relation extraction.

3.2.4.2 Relation extraction

The process of relation extraction aims to identify and capture meaningful connections between entities within BCPSs. As mentioned in the annotation scheme 3.2.2, we have two main relations to extract: the hasSide relation between a tumor entity and a side entity, and the isAttributeOf relation between an attribute and its entity.

The relation extraction procedure consists of the following steps:

Entity identification: After extracting the sections. The NER algorithm introduced in 3.2.3 is employed to extract entities and attributes within each section. Once we have the entities and attributes we can extract the relations between the lesions and their side, and then extract the attributes for the side and the tumor.

hasSide relation extraction: The extraction of the "hasSide" relation focuses on establishing the relationship between tumor entities and side entities. This relation signifies the presence of a tumor on a specific side of the patient's breast, such as left, right, or bilateral. The extraction process involves a rule-based algorithm that identifies tumor entities within each sentence and connects it to the appropriate side.

Side attributes extraction: Side attributes refer to specific characteristics or features associated with the affected sides of the patient. These attributes may include laterality (left, right, or bilateral), BIRADS classification, presence of clinically positive nodes, TNM stage, and other related information. The extraction of side attributes involves a rule-based algorithm that focuses on identifying side-specific entities and extracting their associated attributes. The algorithm utilizes domain-specific knowledge and linguistic patterns to identify and associate attributes with their corresponding side entities, taking into account the context and relationships within the clinical notes.

Lesion attributes extraction: Lesion attributes encompass details and characteristics related to the detected breast lesions. These attributes may include tumor size, tumor grade, presence of associated in situ carcinoma, HER2 status, Ki67 expression level, and other relevant information. The extraction of lesion attributes involves a rule-based algorithm that identifies and associates attributes with their corresponding lesion entity. The algorithm leverages domain-specific knowledge, linguistic patterns, and contextual information to extract the relevant lesion attributes, considering the relationships and context within the BCPSs.

3.2.4.3 Automatic scenario-based data extraction

Cancer care plans are organized around a number of therapeutic modalities such as surgery (SUR), chemotherapy (CHEM), targeted therapies, endocrine therapy (HO), and radiotherapy (RAD). When dealing with non-metastatic breast cancer patients, four periods of interest or "scenarios" can be identified concerning the clinical pathway of the patient:

Scenario A : when cancer has just been diagnosed and no treatment has been performed, the initial therapeutic decision may then be surgery or neoadjuvant therapy;

Scenario B : when a neoadjuvant therapy has been administered;

Scenario C : when neoadjuvant therapy and surgery have been administered;

Scenario D : when only surgery has been first performed and adjuvant treatment modalities have to be decided.

The diagram displayed in figure 3.7 illustrates all the possible trajectories. Depending on the scenario, the information needed to describe the patient's case and to make appropriate decisions varies. For example, if we are in scenario A (initial decision), we would want to know the BIRADS classification score, so that we know the risk of cancer in each breast. However, this information (the BIRADS) would be useless in scenario D because we already did the surgery and we want to know what kind of treatment is recommended after the surgery. So the BIRADS doesn't add value when we are in scenario D.

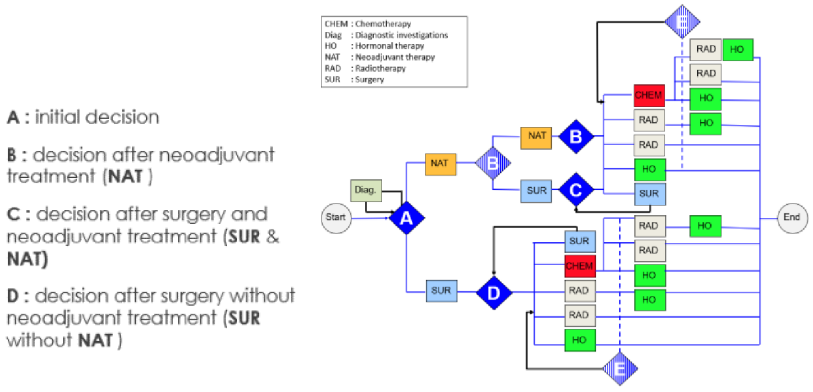


Figure 3.7: Diagram of all medically relevant care plans for non-metastatic breast cancer (Kouz *et al.*, 2020)

3.2.5 Evaluation

To conduct the evaluation, we utilized a corpus of manually annotated breast cancer patient summaries. Each BCPS was meticulously reviewed by a domain expert, who annotated the data using the BRAT annotation tool.

Since we employed BRAT for annotations, we structured the evaluation into three distinct modalities: Entity Mentions, Attribute Mentions, and Relation Mentions. Entity Mentions correspond to mentions of attributes as described in Table 3.2, whereas Attribute Mentions pertain to the values (also described in table 3.2) for the attributes mentioned in the same table. For instance, in the text "Breast classified ACR 4," the extracted entity mention is 'ACR 4,' and the corresponding attribute mention for this entity would be "Birads4."

The final modality, Relation Mentions, encompasses two key relations: "isAttributeOf," representing the relation between a tumor and its attributes, and "HasSide," representing relations between a side and its attributes or lesions.
In concise terms, the evaluation methodology addresses different aspects of the information extraction system:

1. Entity Mentions: We assess the performance of the NER pipeline in accurately extracting mentions of entities and their main attributes. This evaluation showcases how well the system identifies and classifies entities in the text.

2. Attribute Mentions: Here, we evaluate the postprocessing methods' efficiency in extracting attribute values associated with the identified entities. This assessment allows us to gauge the system's ability to capture and extract relevant attribute values from the text. We also use attribute mentions to express context-related entities such as negation.

3. Relation Mentions: The evaluation of Relation Mentions focuses on examining the model's capacity to perform section-based reasoning. This enables the extraction of relations between different entities, allowing us to understand the interactions and associations between them.

To evaluate Entity Mentions, we compiled a list of triplets containing the start position, end position, and the type of mention for each annotated file. We then compared this list of mentions in the gold standard (manually annotated) to the list of mentions predicted by the algorithm. The evaluation was carried out using standard performance metrics: precision, recall, and F1-score. True positives correspond to triplets that appear in both lists, false negatives are triplets present in the gold standard but not in the predicted list, and false positives are triplets found in the predicted list but not in the gold standard.

Similarly, to evaluate Attribute Mentions, we created triplets consisting of the type of attribute, the value of the attribute, and the ID of the corresponding entity mentioned for each annotated file. The performance of Attribute Mentions was evaluated using the same standard metrics: precision, recall, and F1-score.

For Relation Mentions, we generated triplets containing the IDs of the first and second entities involved in the relation, along with the type of the relation. This information allowed us to evaluate the relations using precision, recall, and F1-score.

By conducting this thorough evaluation, we sought to precisely gauge the system's accuracy in identifying and classifying entity mentions, attribute mentions, and relation mentions. The evaluation results provide valuable insights into the strengths and areas for improvement of the structured data extraction pipeline. It allows us to iteratively refine the system, enhance its precision,

and ensure the reliable extraction of crucial information from breast cancer patient summaries. Ultimately, this evaluation process plays a crucial role in building a robust and effective information extraction system for breast cancer patient data.

3.3 Results and discussion

The rule-based algorithm for attributes and values extraction from BCPSs was developed using small BCPS samples, which were iteratively discussed with domain experts to refine the annotation scheme and add new rules for capturing novel entities. After six manual annotation sessions with 2 advanced experts, we identified all the relevant attributes and entities to extract. The algorithm underwent evaluation on a manually annotated corpus of 30 BCPSs.

The evaluation results, presented in Table 3.4, demonstrate the algorithm's efficiency in extracting entity mentions corresponding to various attributes described in Table 3.2. For attribute values extraction requiring post-processing, we utilized BRAT's attribute mention feature. In Table 3.4, the performance of values extraction for attributes is highlighted in bold.

Table 3.4: Pipeline's performance for attributes and values

	precision	recall	F1-score	Number of mentions
Patient entity				
BraSize Cup	0.81	0.76	0.78	26
Breast cancer relapse	0.80	0.73	0.76	7
Comorbidities	0.91	0.64	0.72	86
Comorbidities values	**0.81**	**0.77**	**0.78**	**45**
Drugs	0.79	0.82	0.78	68
Menopausal status	0.94	0.94	0.94	20
Menopausal status values	**0.94**	**0.94**	**0.94**	**20**
Response assessment to neoadjuvant therapy	0.84	0.78	0.79	22
Response assessment to neoadjuvant therapy values	**0.84**	**0.78**	**0.79**	**22**
Genetic mutation	1.00	1.00	1.00	6
Genetic mutation values	**0.83**	**0.83**	**0.83**	**6**
OncotypeDX	1.00	1.00	1.00	3
Side entity				
Side	0.97	0.97	0.97	520
BIRADS classification	0.96	0.95	0.96	126
BIRADS value	0.96	0.95	0.96	126
Confirmed Positive Nodes	0.29	0.29	0.29	20
Confirmed Positive Nodes values	**0.29**	**0.29**	**0.29**	**20**
Cavity Shave Margin	0.91	0.85	0.87	26
Cavity Shave Margin values	**0.98**	**0.93**	**0.94**	**33**
Clear Surgical Margins	0.64	0.60	0.62	58
Clear Surgical Margins values	**0.64**	**0.60**	**0.62**	**58**
TNM	0.90	0.82	0.85	48
N status	1.00	1.00	1.00	17

Clinical Positive Nodes	0.80	0.76	0.77	123
Clinical Positive Nodes values	**0.81**	**0.75**	**0.76**	**125**
Widespread Microcalcifications	0.82	0.82	0.82	21
Widespread Microcalcifications values	**0.82**	**0.82**	**0.82**	**21**
NodeSize	0.85	0.73	0.75	15
Lesion entity				
Estrogen receptor	0.96	0.96	0.96	62
Progesterone receptor	0.96	0.96	0.96	61
HER2 status	0.96	0.95	0.96	64
Ki67	0.96	0.95	0.95	53
Tumor size	0.93	0.90	0.91	167
Tumor grade inv	0.94	0.94	0.93	57
Tumor grade inv values	**0.94**	**0.95**	**0.94**	**57**
Tumor grade insitu	1.00	1.00	1.00	23
Tumor grade insitu values	**1.00**	**1.00**	**1.00**	**23**
Tumor site	0.92	0.93	0.92	145
Tumor site values	**0.85**	**0.93**	**0.88**	**145**
Histology	0.92	0.87	0.88	119
Histology values	**0.90**	**0.85**	**0.87**	**119**
Tumour	0.93	0.92	0.92	317
Presence Emboly	1.00	1.00	1.00	19
Presence Emboly values	1.00	1.00	1.00	20
FISH	1.00	0.86	0.90	13
Associated InSitu Carcinoma	1.00	1.00	1.00	14
AssociatedInSituCarcinoma values	**1.00**	**1.00**	**1.00**	**14**
Diagnosis procedures				
Biopsy	0.95	0.93	0.94	68
Biopsy values	**0.96**	**0.95**	**0.95**	**46**
Cytoponction	0.92	0.92	0.92	18
Cytoponction value	1.00	1.00	1.00	1
Ultra sound	1.00	1.00	1.00	55
MRI	0.95	0.95	0.95	48
Mammography	0.93	0.90	0.91	46
Pet scan	0.92	0.89	0.90	46
Clinical examination	0.96	0.96	0.96	49
Treatment procedures				
Anti HER2 therapy	0.37	0.37	0.37	12
Anti HER2 therapy values	0.00	0.00	0.00	**7**
Surgery	0.92	0.83	0.87	220
Surgery values	**0.99**	**0.91**	**0.94**	**211**
Radiotherapy	0.95	0.87	0.90	37
Radiotherapy values	**0.75**	**0.54**	**0.62**	**7**
EndocrineTherapy	0.83	0.75	0.77	22
Chemotherapy	0.83	0.81	0.81	59

The evaluation was carried out on attribute and value extraction for different entities, including "Patient," "Side," "Lesion," "Diagnosis procedures," and "Treatment procedures." Within each entity, attributes and values were grouped based on the number of mentions they received for a more comprehensive analysis. In the next paragraphs, we will focus on the main entities (patient, side, and lesion).

In the "Patient entity," results indicate relatively high performance for most attributes, with precision ranging from 0.79 to 1.00, recall from 0.64 to 1.00, and F1-score from 0.72 to 1.00. Notably, Genetic mutation and OncotypeDX achieved perfect precision, recall, and F1-score, likely due to their limited occurrences (6 and 3 mentions, respectively). The Menopausal status achieved the highest F1-score of 0.94 (as it is generally well expressed in BCPSs). The performance of ResponseAssessmentToNeoadjuvantTherapy was also notable with an F1-score of 0.79. Some entities, such as BraSize Cup and BreastCancerRelapse, showed slightly lower performance with F1-scores of 0.78 and 0.76. We noticed a poor recall for the comorbidity attribute (0.64) as there were some comorbidities annotated by the experts that were not taken into account when building the rules.

Moving to the "Side entity," the NLP pipeline demonstrated excellent performance in recognizing entities like "N Status", "BIRADS classification," and "Cavity Shave margins" with high F1-scores of 1, 0.97, and 0.87 respectively. Entities such as "WidespreadMicrocalcifications" and "ClinicalPositiveNode" also showed commendable results, with an F1-score of 0.82 and 0.76. Additionally, the entity "TNM" achieved a solid F1-score of 0.85. However, there were some challenges with entities like "Confirmed Positive Nodes" which had the lowest F1-score of 0.29.

Finally, the "Lesion" entity": "Estrogen Receptor," "Progesterone Receptor," "HER2 Status", and "Ki67" attributes exhibited high performance, all achieving precision, recall, and F1-score greater than 0.96. "Tumor Size" and "Tumor Site" attributes also performed well with F1-scores of 0.91 and 0.92, respectively. "FISH" attribute extraction showed a comparatively lower but still very good F1-score of 0.90.

3.3.1 Contextual information extraction

The analysis of the NLP pipeline performance to extract contextual data is described in table 3.5. We can see varying performance across the contextual attributes. Hypothetic showed the best performance, achieving a high F1-score of 0.88, indicating that the model successfully recognized this attribute with good precision and recall. Negation also exhibited notable performance, with an F1-score of 0.74, suggesting accurate identification of this attribute in the text.

However, the attributes Family and Antecedant showed relatively modest results, with F1-scores of 0.63 and 0.51, respectively, indicating room for improvement in their recognition. The family had decent precision and recall, while the antecedent had slightly lower precision and recall values.

On the other hand, the attribute PatientPreference was mentioned 6 times in the expert's annotations and displayed poor performance, with all metrics being 0., indicating that the algorithm struggled to recognize this attribute effectively, in fact, this attribute was only found in 2 BCPSs when building the rules, and these rules did not match any of the mentions of patient preference in the evaluation dataset. This is why it was not included in the table.

Table 3.5: Pipeline performance for contextual information extraction

	precision	recall	F1-score	Number of mentions
Negated	0.88	0.66	0.74	100
Hypothetic	0.91	0.86	0.88	261
Family	0.65	0.62	0.63	26
Antecedant	0.60	0.46	0.51	78

3.3.2 Relation extraction

As explained in section 3.2.4.2, as BRAT makes it complicated to visualize long-distance relations, we expressed the hasSide relation using three attribute mentions: BilateralSide, RightSide, and LeftSide. Table 3.6 shows the performance of the NLP pipeline for the relations extraction, in addition to the hasSide relations, we explore the IsAttributeOf relation between the tumor entity and its attributes.

The analysis reveals that the relationship type IsAttributeOf demonstrated excellent performance, with high precision, recall, and F1-score, all at 0.9, indicating accurate extraction of instances for the relationship between a tumor entity and its attributes. Additionally, the substantial number of mentions (695) further supports the model's strong performance.

Regarding the RightSide and LeftSide relationships, the model demonstrated respectable performance. RightSide had an F1-score of 0.79, and LeftSide performed slightly better with an F1-score of 0.83, suggesting good precision and recall for the extraction of relationships between side entities and their attributes and lesions.

BilateralSide, being a special case in side relation extraction, should be evaluated differently. It typically involves bilateral attributes, and given the complexity of such relationships, and the heuristic chosen and described in section 3.2.4.2, the model achieved modest results with an F1-score of 0.25.

Table 3.6: Pipeline performance for relation extraction

	precision	recall	F1-score	Number of mentions
IsAttributeOf	0.9	0.9	0.9	695
BilateralSide	0.27	0.24	0.25	23
RightSide	0.81	0.79	0.79	350
LeftSide	0.88	0.80	0.83	658

3.3.3 Discussion

In this study, our focus was on assessing the efficiency of a rule-based method for structured data extraction from clinical notes. The overall performance of the pipeline is good, with an average precision and recall of 0.81 and 0.84, respectively. Unsurprisingly, the results revealed that the pipeline performed exceptionally well on attributes where we devoted considerable time and effort in crafting precise rules, for instance, the NLP pipeline had an average F1-score of approximately 0.93 for the Lesion entity attributes which are those for which we put the most effort in the rules as they are important for decision making. This finding underscores the significance of tailoring rules

to specific attributes, leading to improved precision, recall, and F1-scores, thereby enhancing the overall performance of the pipeline.

Interestingly, during the evaluation, we identified the emergence of previously unrecognized attributes, such as "Pregnancy Status" and "AntiHER2 Treatment Value." This discovery highlights the dynamic nature of clinical data, necessitating adaptability in the pipeline to accommodate new and evolving attributes as they surface. Moreover, attributes where we achieved poor performance like "Clear surgical margins " have given us insights on new patterns to represent these attributes, which led to the update of the rules. Typically for the "Clear surgical margins" attribute, we found that the expert set 3mm as the distance for which we consider the margins should be considered as clear (meaning that enough healthy tissue surrounding the cancerous tissue was removed during surgery) whereas it was considered 5mm in the rules created.

On the other hand, certain attributes exhibited suboptimal performance, which could be attributed to their complex expressions that posed challenges for traditional rule-based extraction methods, typically the "Confirmed positive nodes" attribute was very poorly extracted from BCPSs with an F1-score of 0.29. This attribute is often expressed in various language expressions depending on the author of the BCPS. Such findings underscore the need for a more flexible and sophisticated approach, potentially incorporating advanced techniques like deep learning-based models, to handle complex concepts effectively.

Furthermore, the effectiveness of the rule-based method for data extraction is heavily dependent on the quality of the utilized data. BCPSs, presented in a free-text format, exhibits considerable variation in style and content due to the diverse preferences and practices of individual healthcare professionals. This lack of standardized structure poses a significant challenge for rule-based algorithms, leading to inconsistent interpretation and extraction of data. The presence of medical jargon, abbreviations, and context-specific language in clinical notes further introduces ambiguities that can confound the algorithm. Moreover, the absence of clear headers or standardized sections complicates the process of identifying specific information within the notes. Consequently, despite efforts to extract relevant sections and comprehend the text's structure using rules, the algorithm occasionally misinterprets crucial details or fails to capture essential data. Such inaccuracies become particularly problematic if the algorithm is employed to integrate data in a clinical decision support system, potentially leading to errors that could impact patient care and safety. To address these challenges, more advanced NLP techniques, such as machine learning and context-aware language models, should be considered to enhance data extraction accuracy and improve the reliability of CDSS performance. Ongoing efforts by the ANS in France (Agence du Numérique en Santé) and **INCa** (Institut National du Cancer) are in progress to establish such standards for cancer patient summaries, but the widespread adoption of these content guidelines within medical practices takes time and effort.

Additionally, another limitation arose from the relatively small annotated dataset, comprising only 30 BCPS (mainly due to a lack of experts availability). While this dataset was sufficient for the initial evaluation of rules, a larger annotated dataset would have provided more robust results and further insights into the algorithm's performance. However, it should be noted that structured data created manually (in chapter 5 were utilized for additional evaluation, offering an opportunity to validate and complement the findings.

Upon comparing the results to similar works, specifically, the study conducted by Schiappa *et al.* (2022) on breast cancer data extraction using deep learning and keyword methods called RUBY, this pipeline demonstrates comparable and sometimes superior performance.
In terms of the extracted attributes, the NLP pipeline implemented in this work and RUBY showcase close results. For instance, when identifying concepts like "Histologic type," the pipeline

implemented in this work achieves 92% precision compared to RUBY's 81%. Similarly, for the "Menopausal status," the pipeline implemented in this work achieves 94%, outperforming RUBY's 80%. However, RUBY achieves slightly better results for laterality detection with 90%, while the pipeline implemented in this work averages 84%.

Overall, these comparisons highlight that a robust rule-based approach, like the one employed in this NLP pipeline, can achieve competitive results when compared to methods like RUBY which rely on deep learning and keyword techniques. Nonetheless, as we look to the future, a promising avenue for improvement lies in leveraging deep learning methods for context extraction. By annotating more data, the pipeline can capitalize on the power of deep learning models, such as recurrent neural networks and transformers, to capture contextual dependencies and nuances present in clinical notes. This approach is likely to result in enhanced precision and recall, particularly when dealing with complex attribute extractions. However promising results of the rule-based methods also prove that maybe there is no need to go for bigger, black box models.

3.4 Conclusion

In conclusion, this study has demonstrated the efficacy of a rule-based method for structured data extraction from clinical notes. The pipeline achieved remarkable performance on attributes where precise rules were tailored, emphasizing the importance of customization for improved precision, recall, and F1-scores. The discovery of previously unrecognized attributes highlighted the dynamic nature of clinical data, necessitating adaptability in the pipeline to accommodate new and evolving attributes. However, challenges were observed with certain complex attributes, prompting the need for a more flexible and sophisticated approach, potentially involving machine learning-based models.

The quality of the data used in the study also emerged as a critical factor influencing the effectiveness of the rule-based method. The lack of standardized structure and the presence of medical jargon and context-specific language in clinical notes posed challenges for consistent data extraction. Suboptimal performance on some attributes indicated the limitations of traditional rule-based methods and the necessity of advanced NLP techniques.

Looking ahead to adapt to the dynamic nature of clinical data and overcome challenges with complex concepts, a hybrid approach that combines rule-based techniques with deep learning methods holds promise. By embracing this approach and investing in annotated data, the pipeline can evolve into a powerful tool for precise and comprehensive context extraction, further augmenting its utility in clinical settings.

CHAPTER 4

Breast cancer complexity learning

In this chapter, we work on learning the complexity of a breast cancer case from clinical notes. Machine learning (ML) algorithms in natural language processing have been successfully applied for tasks such as the classification of patient record notes, or other documents showing satisfying results. This study aims to compare classical machine learning models with current state-of-the-art language models and rule-based methods on a corpus of annotated real-world breast cancer patient summaries, to identify complex clinical cases in breast cancer patients. The results suggest that classical ML models, specifically multi-layer perceptron, outperformed transformers, pre-trained language models, and the rule-based method, possibly due to the feature extraction based on semantic annotators. However, the study also highlights the limitations of ML and transformer-based models in interpreting outcomes and the need for better structuration of data to express complex clinical concepts. We suggest that further fine-tuning and feature engineering may improve the performance of transformer models, but classical ML models remain a promising approach for this task.

4.1 Introduction

As mentioned in the introduction, patient clinical cases may be of various levels of complexity (Soukup *et al.*, 2019) and there is no a priori definition of breast cancer complexity and very few tools are available that assess cancer complexity (Soukup *et al.*, 2020). In order to delve deeper into understanding and predicting the complexity of patient clinical cases, we explored various approaches, including machine learning and deterministic knowledge-based approaches, using texts or data.

Text classification in healthcare Classification of healthcare texts is considered a special case of text classification. Supervised machine learning algorithms in NLP have been successfully applied. E.g. Support Vector Machines and Latent Dirichlet Allocation have been used for tasks such as classification on patient record notes (Cohen *et al.*, 2014), or other documents in diseases like diabetes

showing satisfying results (Marafino *et al.*, 2014; Wang *et al.*, 2007). These methods required manual feature selection which can be a challenging and time-consuming process, particularly when working with large and complex datasets. manually selecting features such as n-grams or bags of words can also result in a limited representation of the text, as important information and relationships between words may be lost. Additionally, the effectiveness of the chosen features can vary depending on the dataset and require further experimentation and adjustment.

In the past decade, NLP has shown significant advancements, leading to the creation of novel language models like Word2vec (Mikolov *et al.*, 2013), FastText (Bojanowski *et al.*, 2017), and the more recent Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al.*, 2019). These models have significantly reduced the need for manual feature engineering by automatically learning complex text representations. Pre-trained language models like BERT and RoBERTa (Liu *et al.*, 2019b) have been trained on vast amounts of texts, allowing to capture subtle relationships between words and phrases. These models can then be fine-tuned on specific healthcare tasks with a small amount of labeled data, reducing the need for manual feature engineering. They have given good results when fine-tuned on different downstream clinical tasks, including text classification (Li *et al.*, 2022). Furthermore, transformers can adapt to different types of text data, from clinical notes to scientific publications, with minimal modifications. This adaptability makes them particularly useful in healthcare, where the language used can vary significantly depending on the context and specialty.

In this chapter, we have used a corpus of annotated BCPSs to classify them according to their complexity. We considered this problem as a binary classification task, and we compared two different methods to extract features from the texts:

- Using semantic annotators (Sakji *et al.*, 2010; Aronson & Lang, 2010) to extract features from the text as data and use them to train classical machine learning algorithms on BCPS classification,
- Using pre-trained language models and training them on the text of BCPSs for the classification task.

4.2 Material and methods

4.2.1 Data annotation by experts

Between November 2020 and January 2022 (15 months), we asked experts from Tenon Hospital's breast cancer MTBs to routinely annotate every patient discussed either as complex or as non-complex. When deciding a clinical case was complex, experts had to provide the reason for the complexity (e.g., the case is complex because the patient is pregnant). Reasons for complexity were collected to implement a symbolic rule-based method for complexity prediction and to try to approach a local formal definition of case complexity.

4.2.2 Learning complexity using automatic semantic annotators

As we focused on learning complexity before the full effective implementation of structured data extraction in chapter 3, we initially decided to use automatic semantic annotators and indexers to structure the relevant content of natural language BCPSs. This section describes the process of using annotators to extract clinical concepts from the texts and then learning the complexity using this representation.

4.2.2.1 Clinical concepts extraction

We previously used semantic annotators to extract structured data from clinical notes (Redjdal *et al.*, 2022a). Among several semantic annotators, we chose to work with ECMT (Sakji *et al.*, 2010) as it is made for the French language and MetaMap(Aronson & Lang, 2010) because it's widely used for the English language.

ECMT (*Extracteur de Concepts Multi-Terminologique*) is a web service designed for information retrieval in the French language. It takes inspiration from the CISMef algorithm and combines it with the Doc'CISMeF search engine and F-MTI, a multi-terminology automatic indexer. ECMT offers two query modules: a default module that uses a bag of words algorithm and an expanded module that utilizes textual indexing with Oracle text indexing. It employs seven pre-defined terminologies and supports semantic expansion features (Pereira *et al.*, 2008).

MetaMap (Aronson & Lang, 2010) was developed by the National Library of Medicine to map biomedical texts to concepts in the Unified Medical Language System (UMLS). The tool uses a hybrid approach combining natural language processing, a knowledge-intensive approach, and computational linguistic techniques (Aronson, 2001).

To take advantage of MetaMap, we had to automatically translate French BCPSs into English. Since BCPSs are textual documents containing a lot of abbreviations, acronyms, and specialized terms related to the oncology field (e.g., "Echo", "IRM", "TEP"), a first step was to disambiguate the texts. To solve this issue, we created a local dictionary with medical acronyms and their expansion based on online available dictionaries. Then, we replaced acronyms in BCPSs by their expansion and finally used the pre-trained *Opus-MT* translation model (Tiedemann & Thottingal, 2020). As a result, all BCPSs were available in both French and English. We executed the 2 annotators and processed the output of each annotator to generate a semantic representation of a BCPS as two vectors, a vector of UMLS concepts (CUI) extracted with MetaMap, and a second vector containing the labels of the concepts extracted by ECMT (ECMT does not extract UMLS CUIs). For each concept, we associated information about negation as attached to the concept provided by the annotators (e.g., in "absence d'adénopathie", the adenopathy concept was present but identified as negated). Figure 4.1 depicts the whole sequence implemented for the extraction of clinical concepts from BCPSs and the rule-based classification that is described below.

4.2.2.2 Rule-based complexity classification

As explained in section 4.2.1, we asked MTB experts to give the reason for the complexity after discussing clinical cases. To use this information, and after having manually verified that semantic annotators were able to extract medical concepts related to the complexity of a cancer case (Redjdal *et al.*, 2021a), we created a set of rules that match complexity-related concepts if any, and classified BCPSs as complex or not. To get these concepts, we analyzed the justifications provided by the experts to explain the reason for the complexity. The analysis led to the selection of a vector of concepts that represent the "complexity-related concepts". When at least one concept was present in a BCPS, the case was considered complex. Figure 4.1 illustrates the rule-based method used following the concepts extraction.

Furthermore, we did a comprehensive analysis of complexity, drawing from the expert-provided rationales for the cases they classified as complex. The primary objective of this analysis was to try to derive a comprehensive definition of complexity rooted in two years' of clinical data. These insights would serve to deepen our understanding and contribute to the knowledge within the breast cancer domain.

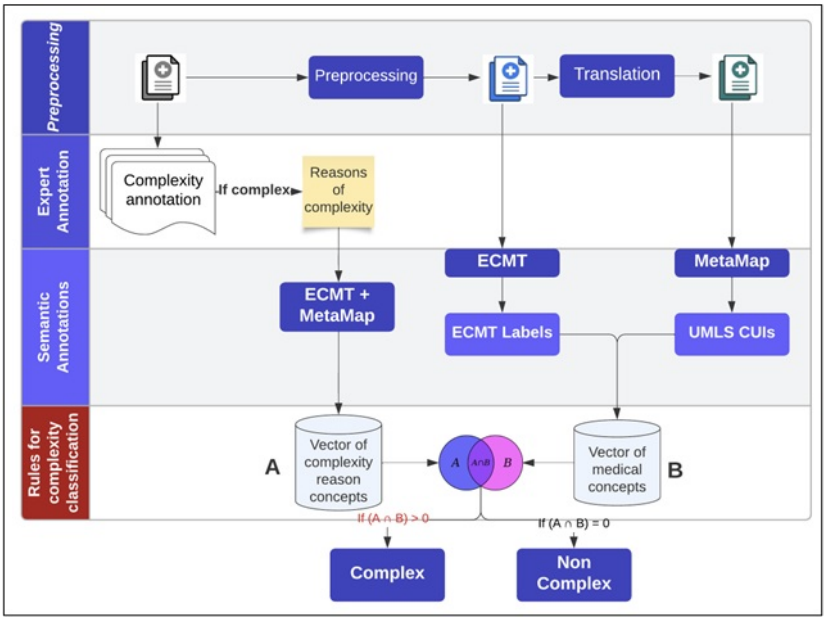


Figure 4.1: Clinical concepts extraction and rule-based classification.

4.2.2.3 Machine learning-based complexity prediction

Data preprocessing In order to get a BCPS representation consumable by all learning models starting from the two vectors obtained from annotators, we converted each BCPS into a row of features that represented the clinical concepts. We included all the labels of the concepts extracted and the value for each feature was 1 if the concept was present, 0 if the concept was not present, and -1 if it was present and negated. We preserved the order of concepts as expressed in a BCPS by using an index column to specify the order in which they appeared in the text.

Model training pipeline We published a work where we tried several machine learning and deep learning model for BCPS complexity classification (Redjdal *et al.*, 2022b). Among these models XGboost (Chen & Guestrin, 2016) and a multi-layer perceptron (MLP) (Popescu *et al.*, 2009) were the best-performing models. In this chapter, once the features were extracted and the data processed, we trained an MLP and an XGboost model on the annotated data. We used a k-fold cross-validation strategy, where the model hyper-parameters tuning process was executed using Grid Search (Liashchynskyi & Liashchynskyi, 2019). The resulting classification models were evaluated using precision, recall, and F1-score. Because the data of this analysis is unbalanced, with less complex cases than non-complex cases, we used both the ROC curve and the Precision-Recall PR curve to evaluate the performance of the binary classification model. While the ROC curve evaluates the trade-off between sensitivity and specificity at different classification thresholds, the PR curve measures the trade-off between precision and recall. This is particularly useful since the cost of false positives and false negatives is not equal in this case. We also calculated the area under the PR curve, which measures the overall performance of the model at different classification thresholds. Figure 4.2 describes the model training pipeline.

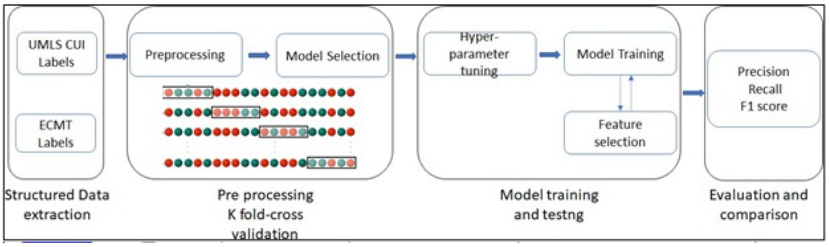


Figure 4.2: Model training pipeline

4.2.3 Learning complexity using pre-trained language models

4.2.3.1 Transformer-based method

We used state-of-the-art language models, based on attention and transformer architecture in order to do the classification. The main problem we faced was that transformer models accept only 512 tokens as input, whereas most of the BCPSs contain more than that with an average of 2200 tokens per document. This is a known problem especially in the clinical domain (Gao *et al.*, 2021b) and there are several ways to deal with it. We used two strategies:

The first strategy was to try the truncation method (Sun *et al.*, 2019). It consists of dropping some of the tokens in order to fit with the 512 tokens limit. We choose to keep the first 25 tokens (the first tokens contain the reason the patient is discussed by the MTB), and the last 485 tokens because the last part of the text contains the most up-to-date information about the patient.

The second strategy consisted of dividing each BCPS into multiple chunks of 512 tokens or less, following the method described by Pappagari *et al.* (2019). Each chunk is then tokenized, and the list of tokenized chunks given as input to the classifier. A classification is then performed for each chunk and finally, the class of each document is calculated according to its chunk class. For this strategy, we used the same BERT model as the one used in the first strategy.

In order to train the models, we used a BERT model that was trained on all the documents of AP-HP data warehouse (Dura *et al.*, 2022a), we will call it BERT-EDS.

4.2.3.2 Static word embeddings

In addition to exploring state-of-the-art pre-trained language models, we also investigated the effectiveness of earlier embedding methods (these methods are explained in section 2.4.1.1.1), we used Word2Vec and GloVe for this classification task.

For Word2Vec (Goldberg & Levy, 2014), we first preprocessed the text data, tokenized it, and removed stopwords and punctuation to create a clean representation of the clinical notes. Subsequently, we trained the Word2Vec model on the preprocessed texts to generate dense word embeddings that capture semantic relationships between medical terms. Each BCPS was then converted into a feature vector by calculating the average word embeddings of the clinical concepts present in the text. Utilizing this feature representation, we employed a multi-layer perceptron classifier for complexity prediction, optimizing its hyperparameters using GridSearch following the method explained in figure 4.2. We chose MLP as it was the best model to perform this task using the semantic annotators (Redjdal *et al.*, 2022b).

For the GloVe embeddings (Pennington *et al.*, 2014b), which offers pre-trained word embeddings that capture semantic relationships similarly to Word2Vec. We loaded the pre-trained GloVe embeddings and used them to generate dense word representations for the preprocessed text data. These embeddings were then utilized to create feature vectors for each BCPS, employing again an MLP classifier.

4.3 Results and discussion

We conducted this study on a sample of 1,048 BCPSs (763 non-complex cases and 285 complex cases), which correspond to all clinical cases discussed by MTB clinicians between November 2020 and January 2022 at Tenon Hospital in Paris (France). The data set was divided into 80% for training and 20% for validation. The models were trained on the training dataset using a 5-fold cross-validation. The validation set was not explored and was used only to evaluate the results. Table 4.1 summarizes the results obtained by each model on the validation set.

Table 4.1: Evaluation of the models on the validation set

Model	**Precision**	**Recall**	**F1 score**	**Accuracy**
Using semantic annotators				
XGboost	0.84	0.83	0.81	0.83
MLP	0.88	0.89	0.88	0.89
Rule-based method	0.66	0.64	0.65	0.64
Using pre-trained language models				
Word2vec	0.76	0.77	0.73	0.77
GloVe	0.72	0.75	0.72	0.75
Truncation + BERT-EDS	0.53	0.72	0.61	0.72
Chunks + BERT-EDS	0.53	0.72	0.61	0.72

Among the semantic annotators, the XGboost and MLP models achieved solid accuracies of 83% for XGboost and 89% for MLP, while the rule-based method performed at 64% accuracy. In the category of pre-trained language models, Word2vec and GloVe attained decent accuracies of 77% and 75%, respectively, whereas Truncation + BERT-EDS and Chunks + BERT-EDS exhibited lower accuracies, both at 72%.

4.3.1 Using semantic annotators

4.3.1.1 Machine learning classification

Feature extraction using semantic annotators resulted in the extraction of 20,682 unique UMLS concepts and ECMT labels. For XGboost and MLP, The area under the PR curve was 0.88 for MLP and 0.79 for XGboost, and the ROC AUC was 0.92 for MLP and 0.88 for XGboost. Figure 4.3 shows both curves for MLP and XGboost, and Figure 4.4 shows the confusion matrix for each model on the validation set with its best parameters and the optimal threshold.

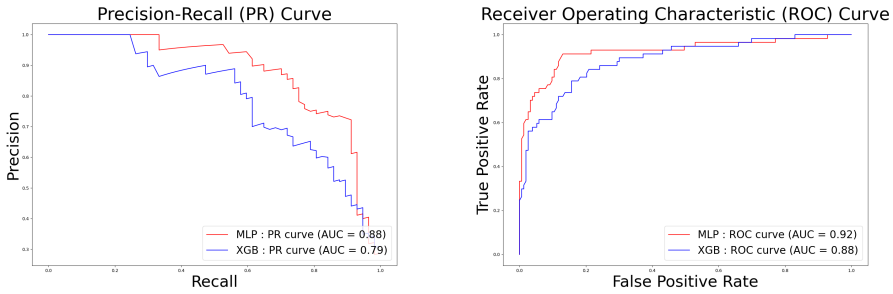


Figure 4.3: PR AUC and ROC AUC for classical ML methods using semantic annotators

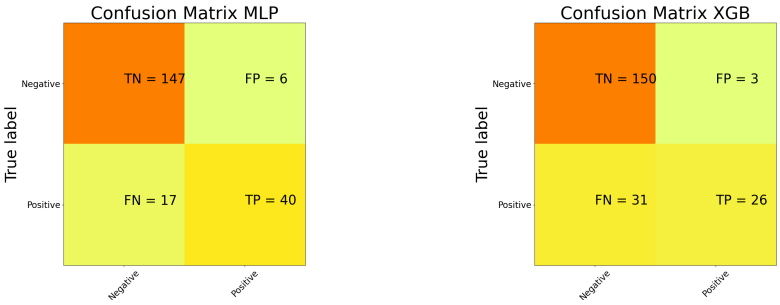


Figure 4.4: Confusion matrices for classical ML methods using semantic annotators

4.3.1.2 Rule-based classification

From the annotated BCPSs, 285 were identified as complex patients. We selected 24 sets of clinical concepts that we deemed as complexity reason concepts. Figure 4.5. presents the results of the rule-based method applied to the validation data, we can see that this method gets a significant number of false positives (44), and performs worse than ML-based methods to extract complex cases (26 true positives against 40 for MLP). The rule-based method's overall performance was lower compared to machine learning methods. It achieved a precision of 0.66, a recall of 0.64, an F1 score of 0.65, and an accuracy of 64%.

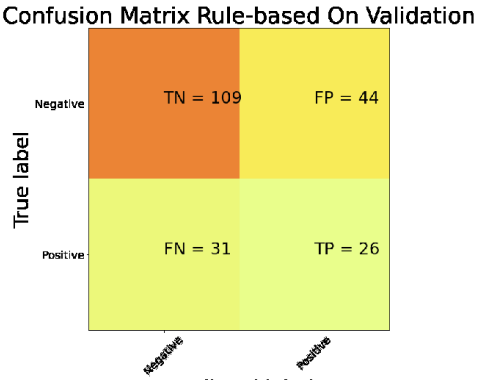


Figure 4.5: Confusion matrice for the rule-based method

Complexity definition The analysis of complexity factors provided by domain experts has resulted in the identification of key characteristics that shed light on complex cases. Following a thorough discussion with the experts to validate this analysis, the findings are summarized below. This summary serves as the definitive definition of a complex case for this study and has been reviewed and

approved by the domain experts.

Histology – Related Complexity

- Oligometastatic cancer
- Breast sarcoma
- Solid or encapsulated papillary carcinoma
- Low-grade adenosquamous carcinoma
- Low-grade metaplastic carcinoma
- Lymphomas
- Phyllodes tumors
- Bilateral breast cancer with different histologies

Complexity due to Comorbidities

- Heart failure or cardiac fragility (contraindication to anthracyclines and sometimes radiotherapy)
- Porphyria (contraindication to certain chemotherapy regimens)
- Scleroderma, Xeroderma Pigmentosum (contraindication to radiotherapy)
- Associated hematological pathology (e.g. dysmyelopia) increasing the risk of hematological toxicity of chemotherapy
- Associated hematological pathology (e.g. dysmyelopia) increasing the risk of hematological toxicity of chemotherapy
- Hepatocellular insufficiency
- Active autoimmune disease and immunotherapy
- Renal insufficiency

Complexity for Other Reasons (Non-standard Treatment Procedure)

- Pregnancy (some treatments cannot be administered: anti-HER2, immunotherapy, antihormonal treatments)
- Non-operable patients (for example, old patients)
- Presence of another cancer with the breast cancer
- Male breast cancer
- Presence of BRCA mutation
- Eligibility criteria in clinical trial/eligibility for OncoType (Syed, 2020)
- Not enough information available to classify the patient on a standard treatment
- Patient refusal of standard treatment

4.3.2 Using pre-trained models

Regarding the pre-trained language models, both Word2vec and GloVe models demonstrated competitive performance and were close in their effectiveness in detecting complex cases. Specifically, the Word2vec model achieved an accuracy of 0.76, a precision of 0.77, a recall of 0.73, and an F1-score of 0.77, while the GloVe model achieved an accuracy of 0.72, a precision of 0.75, a recall of 0.72, and an F1-score of 0.75.

When compared to Word2vec and GloVe, the BERT-EDS methods (Truncation + BERT-EDS and Chunks + BERT-EDS) showed inferior performance, achieving an accuracy of 0.53, a precision of 0.72, a recall of 0.61, and an F1-score of 0.72. These lower scores across all evaluation metrics indicate that BERT was less successful. When looking at the results, both BERT methods classified all the validation data as non-complex. As we can see in figure 4.6 and figure 4.7 both GloVe and Word2Vec also struggle to find complex cases as the number of true positives is 15/58 for Word2Vec and 16 for GloVe), compared to the methods based on semantic annotators and ML algorithms (up to 40 true positives). We also notice that the PR-AUC is lower than machine learning methods with 0.52 for Word2Vec against 0.88 for MLP.

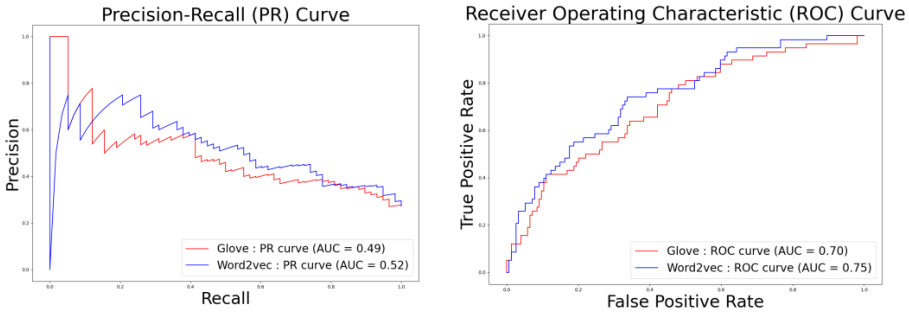


Figure 4.6: PR AUC and ROC AUC for classical ML methods using GloVe and Word2vec

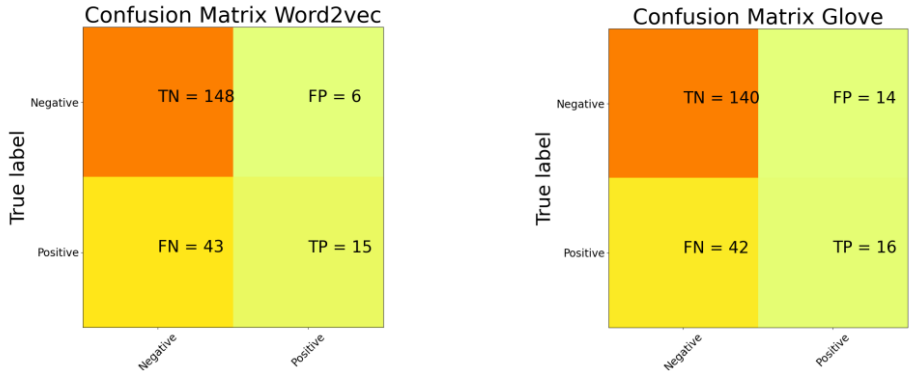


Figure 4.7: Confusion matrices for classical ML methods using GloVe and Word2Vec

4.3.3 Discussion

The findings of our study indicate that classical machine learning (ML) models outperform transformer models in the task of breast cancer case complexity classification. Particularly, the Multilayer Perceptron model exhibited higher performance than XGBoost, especially in detecting complex cases. From the validation set containing 57 complex cases, MLP successfully identified 40 cases, whereas XGBoost only detected 26 true positives. The ROC curve analysis further demonstrated that MLP achieved a higher true positive rate with a lower false positive rate compared to XGBoost, as evidenced by the AUC values of 0.92 and 0.88, respectively. Similarly, the PR curve analysis revealed that MLP outperformed XGBoost, with AUC values of 0.88 and 0.79, respectively, which explains why MLP correctly classified a greater number of complex cases compared to XGBoost.

Among the transformer-based models, both the Truncation + BERT-EDS and the Chunks + BERT-EDS strategy strategy achieved an accuracy score of 0.72. However, relying solely on accuracy may not offer a comprehensive understanding of the models' performance. BERT, in particular, encountered challenges in predicting the positive class (complex cases) and mistakenly classified the entire validation set as non-complex.

This subpar performance could be attributed to data imbalance and the limited training data, which resulted in poor generalization. These findings are in line with previous hypotheses (Gao *et al.*, 2021a) that suggest a pre-trained BERT model may not be the most optimal choice for clinical text classification, given that clinical documents often exceed the standard BERT token limit of 512 tokens. To address this limitation, recent research has proposed new pretraining methods, one of which encourages BERT to learn about entities rather than generic syntax and grammar patterns (Xiong *et al.*, 2020). Such approaches show promise in enhancing the performance of clinical and biomedical classification tasks that require knowledge-based reasoning. Additionally, other studies (Beltagy *et al.*, 2020b) have adapted BERT to handle long texts without the need for hierarchical splitting methods, potentially allowing the model to identify meaningful patterns over longer distances and improve overall performance.

Comparing the performance of Word2vec and GloVe to transformer models, specifically BERT-EDS methods. Classical word embedding models (Word2vec and GloVe) exhibited higher accuracy and outperformed BERT-EDS in correctly identifying complex cases. The effectiveness of Word2vec and GloVe models in detecting complex cases can be attributed to their ability to capture semantic relationships and contextual information within the clinical text data. Utilizing word-level embeddings, Word2vec and GloVe capture word associations based on co-occurrence patterns in large text corpora, allowing them to effectively represent medical terminologies and concepts for accurate complexity classification.

Our findings also highlight that a rule-based method is less effective than classical ML methods for complexity classification. Nevertheless, it is essential to acknowledge that the rule-based method's limitations stem from the fact that the clinical concepts used do not encompass all complexity reasons. Notably, reasons related to patient preferences or contextual data could not be adequately expressed as UMLS or MeSH concepts, leading to incomplete coverage of complexity factors. As a consequence, false positives were present in the rule-based method, as a single reason might not be sufficient to accurately classify a clinical case as complex.
The comprehensive definition of complexity that we developed through collaboration with domain experts represents a valuable foundation for creating a robust rule-based method for complexity prediction. However, to fully leverage the complexity definition's potential, improved data structuring and additional information beyond a simple vector of semantic annotations are re-

quired. Consequently, the structured data extracted in Chapter 3 can play a crucial role in creating rules based on the complexity definition, in close collaboration with experts. This integrated approach has the potential to enhance the rule-based method's performance and improve the accuracy of complexity classification for breast cancer cases. By combining structured data and domain expertise, we may enhance the rule-based approach to encompass a wider range of complexity factors, such as patient preferences and contextual information. Furthermore, the structured data extracted in chapter 3 can serve as input for a machine learning algorithm used in the classification task. Although we began exploring this avenue, we were unable to complete the work due to time constraints.

4.4 Conclusion

Overall, the results suggest that classical ML models currently outperform transformer models and rule-based methods for breast cancer complexity classification. However, it is essential to consider that transformer models with further fine-tuning and feature engineering, might be capable of achieving better results, especially with a larger set of annotated data. Notably, the challenge of interpreting model outcomes arises due to the large number of concepts annotated for each clinical case, hindering the extraction of relevant features and the ability to explain the classification decisions. Future research efforts may focus on addressing this limitation to improve interpretability and provide valuable insights for clinical decision-making.

CHAPTER 5

Update of the guideline-based decision support

With the development of a structured data extraction pipeline and an algorithm that predicts the complexity of clinical cases, we can now proceed with the implementation of decision support modules. This chapter focuses on the guideline-based CDSS for non-complex cases. As explained in the introduction 1.3, we hypothesize that guideline-based CDSSs are useful for dealing with non-complex cases. So for them, the idea is to provide a guideline-based decision support system. To achieve this, we utilize the GL-DSS of the DESIREE project, which is based on the CPGs of AP-HP for breast cancer published in 2016. Since a new version of AP-HP guidelines named SENORIF "Cancers et pathologies du sein attitudes diagnostiques et thérapeutiques, protocoles de traitement" has been published in 2021, the aim of the work in this chapter is to develop a semi-automated method for identifying practice evolutions to update the GL-DSSs knowledge base with the latest evidence. To do this, we use real-world data from the corpus of the BCPSs manually structured and used in section 3.2.5, and we run the GL-DSS on the non-complex cases identified within this corpus. The objective is to compare the decisions made during MTBs to the recommendations produced by the GL-DSS for each clinical case. When MTB decisions does not align with the GL-DSS's recommendation, we refer to SENORIF to assess the MTB decision's compliance with the most recent guidelines. This method permits us to accurately identify patients' profiles for which medical practice has evolved, thus identifying the updates to make to the GL-DSS's knowledge base. Using a sample of 160 patients extracted from the complexity learning dataset and representing non-complex cases, this approach enabled the identification of 38 patients for whom there have been practice evolutions. These 38 profiles involve the addition of 21 rules, modification of 18 rules, and removal of 9 rules from the knowledge base. The addition of these rules was followed by an update of the system's ontology as we added new concepts that did not exist in the initial ontology. Among these modifications, 23 pertain to surgical modalities, 20 to chemotherapy, 4 to targeted therapies, and 1 to radiotherapy.

5.1 Introduction

As discussed in section 2.1.1 CDSSs can enhance the quality of care by promoting the application of CPGs to individual patients and facilitating the utilization of up-to-date clinical evidence, which is crucial for minimizing errors (Pelayo *et al.*, 2020; Voigt & Trautwein, 2023; Mazo *et al.*, 2020; Ricci-Cabello *et al.*, 2023). The formalization of CPGs and their integration into CDSSs also help mitigate non-compliance with recommended practices during decision-making (Mazo *et al.*, 2020).

The acceptance of CDSSs by healthcare professionals plays a crucial role in their routine utilization. Factors such as effectiveness, ease of use, and user-friendly interfaces significantly influence adoption. Additionally, seamless interoperability with EHRs, eliminating the need for redundant data entry, is equally important for CDSS acceptance (Voigt & Trautwein, 2023).

As medical knowledge continues to evolve, it is crucial to regularly review and update computerized CPGs. For guideline-based CDSSs to maintain reliable performance, they require access to the latest evidence-based recommendations. The process of incorporating new knowledge in the form of rules into CDSSs has been implemented in some hospitals. However, this process is often costly and time-consuming (Cánovas-Segura *et al.*, 2019). Currently, there is no fully satisfactory automatic approach to compare two or more CPGs. While the extraction of concepts from CPGs can be efficiently performed using natural language processing methods and standard medical terminologies, the extraction of rules and recommendations still heavily relies on human expertise (Azarpira *et al.*, 2022) (see section 2.2 for more details).

Other approaches have been introduced for the update of knowledge bases, (Bouaud *et al.*, 2007) proposed a method to automatically compare 2 structured CPGs, they consider " CPGs as a set of recommendations R i i. Each recommendation R i is characterized by a pair (Si, Ti) with Si -> Ti, denoting that a treatment plan Ti is recommended in the clinical situation Si". Figure 5.1 resumes the knowledge base modifications that result from CPGs updating.

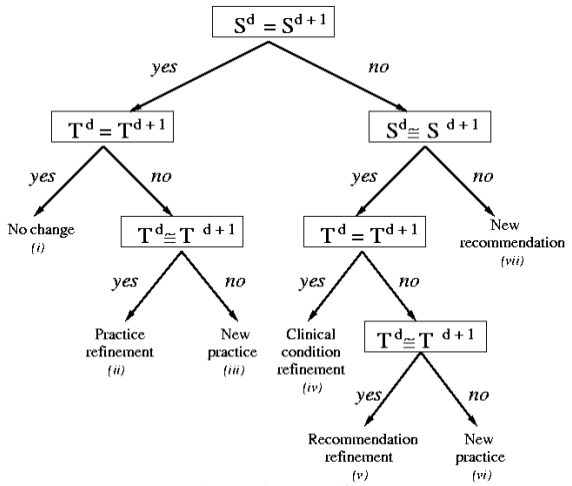


Figure 5.1: Typology of knowledge base modifications resulting from CPGs updating (Bouaud *et al.*, 2007), *S^d: Clinical situation at an initial date; S^{d+1}: clinical situation at new date; T^d: treatment proposed at the initial date*

"Basically 4 main situations are observed by Bouaud *et al.* (2007):

- No change: An identical clinical situation leads to an identical treatment plan.

- Refinement of an existing recommendation. The new recommendation shares some common parts with the former one. Refinement may concern the description of the treatment plan but not the situation, only the description of the clinical situation, or both descriptions.

- New practice. A totally new therapy (noncomparable) appeared in an already identified clinical situation.

- New recommendation: extending the CPG coverage, a new clinical situation is identified with its corresponding therapy leading to a new recommendation."

As stated earlier, our project relies on the GL-DSS system developed for the DESIREE project as a guideline-based Clinical Decision Support System (see section 2.2.4.1. The GL-DSS was initially designed using the **2016 AP-HP** guidelines for breast cancer treatment. However, for the system to be effective, it is essential to update the knowledge base with the most recent evidence available.

SENORIF is a French CPGs for the management of breast cancer. It was developed by a multidisciplinary group of experts in breast cancer from AP-HP, Institut Curie, and Institut Gustave Roussy. Guidelines were last updated in 2021. SENORIF guidelines cover a wide range of topics related to breast cancer, including diagnosis and staging, and treatment options. The guidelines also include several decision-making tools, such as algorithms and flowcharts, to help clinicians make informed decisions about the best treatment for each patient.

In this chapter, we describe the methods used to update the knowledge base of the GL-DSS system developed for the DESIREE project. We rely on real use cases that have been treated in the MTB of the Tenon hospital during the year 2021, and compare the decisions made by the MTB, to the recommendations of the GL-DSS (based on CPGs published in 2016) and the recommendations of SENORIF (2021) for each patient profile. The objective is to capture the knowledge evolution by focusing on a sample of real clinical profiles discussed by Tenon Hospital breast cancer MTB. Knowledge evolution can be expressed in the four situations identified by Bouaud *et al.* (2007): no change, refinement of an existing recommendation, new practice or new recommendation.

We hypothesize that, for non-complex cases, MTB decisions would be compliant with the latest evidence (i.e. SENORF), so profiles, where the decision of the MTB was non-compliant with the GL-DSS recommendations and compliant with SENORIF, were considered as profiles for which there might be knowledge evolution. These profiles were used to update the knowledge base of the GL-DSS by modifying rules, adding new rules and concepts, or removing rules.

5.2 Material and methods

5.2.1 BCKM ontology and GL-DSS inference engine

As mentioned in the introduction, we utilize the Guideline-Based Decision Support System (GL-DSS) of DESIREE to provide guideline-based decision support for non-complex cases. The GL-DSS and the Breast Cancer Knowledge Model ontology are comprehensively described in Sections 2.2.4.1 and 2.2.4.

To evaluate this work, we first installed the GL-DSS in the Jupyter environment of the AP-HP data warehouse. Since we encountered difficulties in installing the graphical user interface (GUI) of the system, we utilized the GL-DSS through the command line interface by invoking the inference engine on the coded patient cases automatically created in the system via the pipeline implemented in chapter 3.

5.2.1.1 Clinical Practice Guidelines

In this chapter, we introduce two guidelines as we aim to update the knowledge base of the GL-DSS. The first is the AP-HP guidelines, which are the guidelines used in the DESIREE GL-DSS. The second one is the SENORIF guidelines, which represent the most recent guidelines on the management of breast cancer patients.

AP-HP CPGs developed by the breast working group of AP-HP, offer therapeutic recommendations for breast cancer discussed in multidisciplinary meetings. It draws upon the expertise of the group, scientific knowledge, professional recommendations, and international guidelines published until 2016. The guidelines consist of 36 pages of narrative text and have been formalized into rules for implementation in the GL-DSS.

SENORIF CPGs presents evolving standards of care in senology based on national and international recommendations, including the French National Cancer Institute and the French National Authority for Health (HAS) as primary references. They encompass screening to advanced disease treatment while excluding non-carcinomatous lesions. Certain options in the guidelines are based on expert consensus when there is a lack of strong scientific evidence, and this is explicitly mentioned. The 2021 version incorporates the latest literature from 2015-2021 and features 188 pages. It includes helpful diagrams and decision trees that summarize the textual content, providing visual aids for practitioners.

5.2.2 Proposed method

To use the GL-DSS of the DESIREE project, and to be able to assess the performance of the GL-DSS, we compared the outputs of the system with the decision made by MTB clinicians on a sample of non-complex BCPSs discussed in 2021. Since these decisions were made on clinical cases discussed in 2021, we had to update the GL-DSS knowledge base to implement 2021 SENORIF CPGs instead of 2016 AP-HP CPGs.

The process was the following: We first transformed the BCPSs of a corpus of non-complex clinical cases into the structured data model of the BCKM, and we then used the GL-DSS to obtain recommendations for these patients according to the 2016 AP-HP guidelines. Then, when one of the recommendations issued by the GL-DSS matched the MTB decision, we considered there was no evolution of practice for that case. If not, we checked the recommendations of the latest French guidelines SENORIF 2021 on the same case. When one of the recommendations manually retrieved from the 2021 textual SENORIF CGPs matched the MTB decision, we considered there was an evolution of practices for this case. When none of the recommendations manually retrieved from the 2021 textual SENORIF CGPs was matching the MTB decision, the case was reviewed by domain experts to check whether the case should be re-classified as complex. Two strong assumptions are made here. The first one is that MTB decisions are expected to be made according to the latest guidelines, i.e. SENORIF. The second one, deduced from the first one, is that SENORIF includes the AP-HP guidelines (as all of the authors of SENORIF were also authors of the 2016 AP-HP guidelines). Figure 5.2 shows the whole pipeline. We consider the following:

- D^{MTB} refers to the MTB decision.
- $R^{GL\text{-}DSS}$ refers to the GL-DSS recommendations based on the computerized 2016 APHP CPGs.

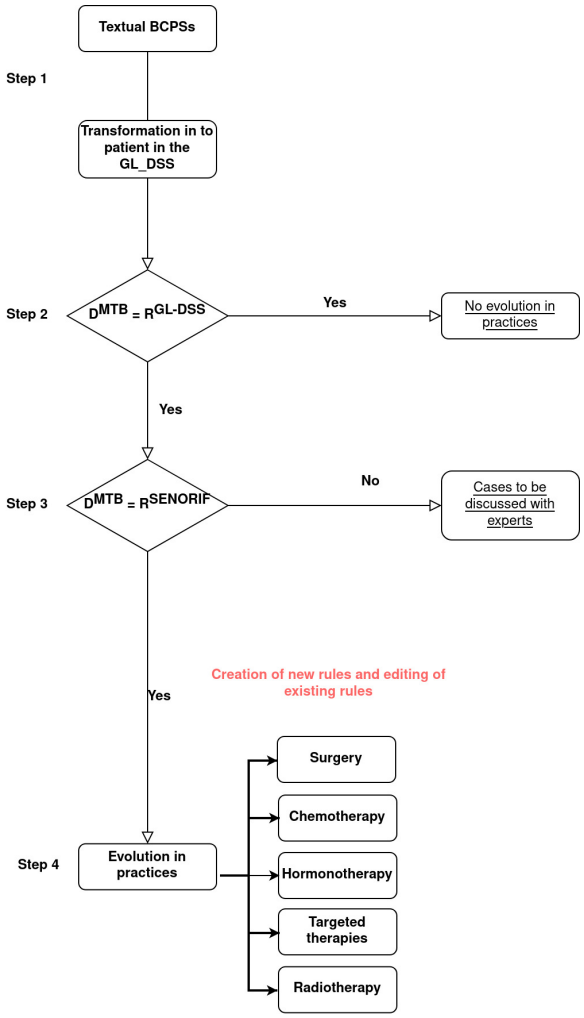


Figure 5.2: Pipeline used for updating GL-DSS's knowledge base

- $R^{SENORIF}$ refers to the SENORIF recommendations manually retrieved from the textual 2021 SENORIF CPGs.

5.2.2.1 Step 1: From BRAT-structured data to patient in the GL-DSS system

Having obtained patient data in a structured format in chapter 3, we proceeded with the mapping process using the BCKM ontology (the ontology is detailed in section 2.2.4). In the BCKM, the number of concepts far exceeds those extracted by the method implemented for structured data extraction. However, it is important to note that a significant portion of the ontology consists of optional data, and only a subset of concepts is utilized for reasoning purposes.

During the mapping process, certain entities in the structured data scheme (described in table 3.2) may have multiple corresponding entities in the BCKM ontology. For example, the data include an attribute called "tumor_size," while the BCKM ontology may define concepts such as "tumor size at MRI" or "tumor size at ultrasound..." To handle this ambiguity, we incorporated the extracted sections from the structured data extraction methodology (see section 3.2.4.1). By analyzing these sections, we guide the mapping algorithm to determine the appropriate concept in

the BCKM ontology. For instance, if the tumor size attribute is found within an MRI section, it can be mapped to the concept of "tumor size at MRI."
Additionally, some entities, such as the TNM classification, require postprocessing to align with the relevant concepts in the BCKM ontology. For example, the annotation scheme has a general attribute for TNM, while the BCKM ontology specifies separate concepts for cT (clinical T of TNM), ycT (residual T assessed by examination or imaging after neo-adjuvant treatment), pT (T of TNM established from the pathology analysis after surgery), and so on. To address this, we employ a postprocessing step that applies predefined rules or heuristics to differentiate and match the components of the TNM classification to their respective concepts in the BCKM ontology. This ensures accurate mapping of the TNM classification from the structured data to the appropriate concepts in the ontology. The same logic is used with other attributes like HER2 status. A table with the mapping of all concepts is provided in the appendix section (Appendix C).
In summary, the mapping process involves connecting the structured data extracted from textual BCPSs as described in chapter 3, with the BCKM ontology, allowing the creation of patient profiles in the BCKM format.

For the rest of this chapter, we use BCPSs of patients treated during the year 2021 at Tenon Hospital. These BCPSs were structured using the algorithm developed in chapter 3, They were then manually created in the GL-DSS by a master's intern who verified the output of the structured data extraction algorithm before entering the patient's information on the GL-DSS. We use this manual verification to make sure that these patients are well entered into the system, avoiding the errors still made by the automatic structured data extraction algorithm.

5.2.2.2 Step 2: Comparison between MTB decisions and AP-HP recommendations produced by the GL-DSS

For each BCPS, we conducted a comparative analysis between the decisions made by MTB clinicians and the recommendations provided by the GL-DSS. The output of the comparison process falls within the following set of categories:

$\mathbf{D}^{\text{MTB}} \in \mathbf{R}^{\text{GL-DSS}}$.

$\mathbf{D}^{\text{MTB}} \neq \mathbf{R}^{\text{GL-DSS}}$ regarding surgery.

$\mathbf{D}^{\text{MTB}} \neq \mathbf{R}^{\text{GL-DSS}}$ regarding chemotherapy.

$\mathbf{D}^{\text{MTB}} \neq \mathbf{R}^{\text{GL-DSS}}$ regarding hormone therapy.

$\mathbf{D}^{\text{MTB}} \neq \mathbf{R}^{\text{GL-DSS}}$ regarding targeted therapy.

$\mathbf{D}^{\text{MTB}} \neq \mathbf{R}^{\text{GL-DSS}}$ regarding radiotherapy.

As a result and according to our assumptions, if we have $\mathbf{D}^{\text{MTB}} \in \mathbf{R}^{\text{GL-DSS}}$, then we consider there is no evolution of practices. However, when the recommendations of the system do not follow the MTB's decision, we go to the next step and try to identify whether it might be a knowledge evolution or not. For example, if we have a patient who is in scenario B or C (i.e. she already had a neoadjuvant therapy, see section 3.2.4.3). and has a positive HER2 status with a non-complete response to the neoadjuvant treatment, the MTB may decide she will benefit from adjuvant chemotherapy, replacing Trastuzumab with TDM1 drug., which is a new practice, while the GL-DSS recommendation is to continue with Trastuzumab. Figure 5.3 shows the difference between the MTB decision and the recommendation.

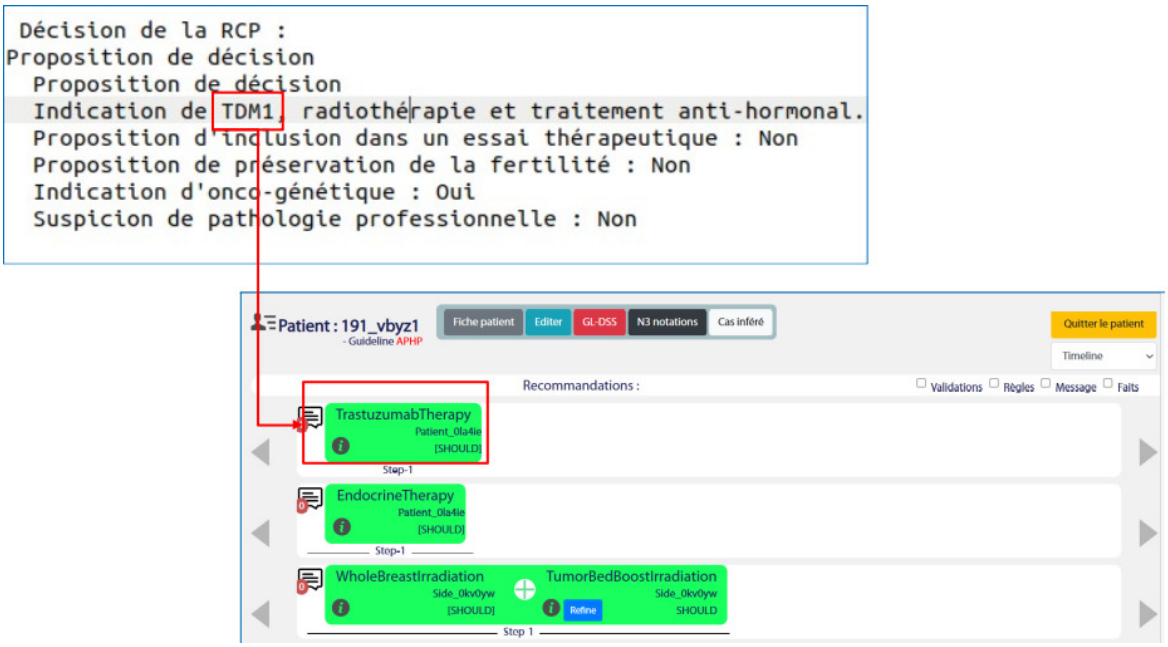


Figure 5.3: Comparing GL-DSS recommendations to MTB decision

5.2.2.3 Step 3: Comparison between MTB decision and SENORIF recommendations

Following the previous step, we selected the BCPSs for which MTB decisions did not comply with the GL-DSS recommendations based on AP-HP CPGs. We then manually examined the SENORIF guidelines to identify which recommendations would apply to these patient profiles and compared them with the MTB decision. Two scenarios could arise:

- $\mathbf{D}^{\text{MTB}} \in \mathbf{R}^{\text{SENORIF}}$: In these cases, MTB decisions align with the SENORIF recommendations but deviate from the GL-DSS recommendation. We consider these patient profiles as instances where there is an evolution of medical practices.
- $\mathbf{D}^{\text{MTB}} \notin \mathbf{R}^{\text{SENORIF}}$: Here, MTB decisions do not align with either the SENORIF or GL-DSS recommendations. These are cases that require further discussion with experts.

For example, continuing with the clinical case presented in the previous section, we refer to the SENORIF guidelines (page 97) and find that the recommendation for these profiles is to replace Trastuzumab with TDM1. We observe in figure 5.4 that this recommendation aligns with the MTB's decision, thus confirming a change in medical practices.

5.2.2.4 Step 4: Identification of updates and creation of new rules

Once we had identified all the profiles for which MTB decisions differed from the recommendations of the GL-DSS, we examined the different scenarios. We confirmed, on the one hand, the evolution of medical practices (when MTB decision $\notin$ GL-DSS recommendations and MTB decision $\in$ SENORIF's recommendations), and on the other hand, we studied the cases where we had MTB decision $\notin$ GL-DSS recommendations and MTB decision $\notin$ SENORIF's recommendations). These cases represent different profiles where the MTB decided not to follow these guidelines.

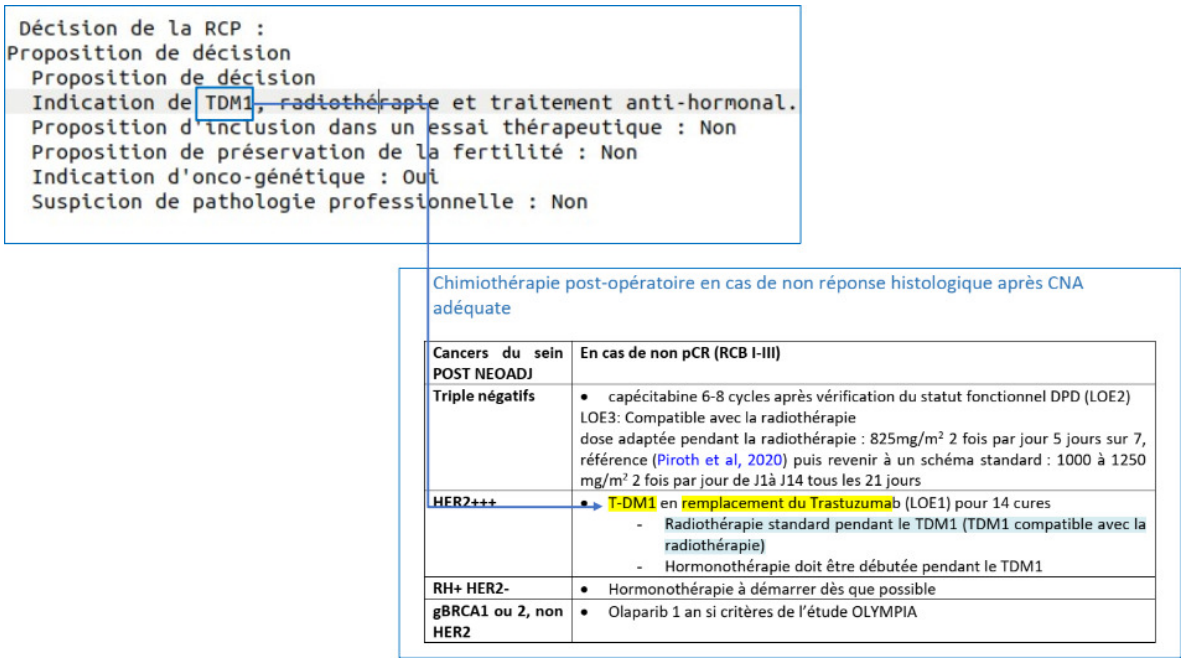


Figure 5.4: Comparing SENORIF recommendation to MTB decision

One possible reason is that the clinical case is not covered by the guidelines and can be considered complex. Additionally, for some cases where the GL-DSS produced incorrect recommendations due to defaults (that were subsequently corrected in the knowledge base).

To update the knowledge base of the GL-DSS, we distinguished between the following two scenarios:

Evolution of practices: In these cases, we proposed to (i) modify existing rules, (ii) add new rules, or (iii) remove outdated rules.

Correction and completion of bugs: Here, the task involved identifying rules in the GL-DSS that were not functioning correctly and required correction. We also added missing rules independent of knowledge evolution.

Beyond modifying the rules in the knowledge base of the GL-DSS using the concepts already present in the BCKM, it was also necessary to update the BCKM by introducing new concepts. In the SENORIF CPGs, there are indications for new treatments and recommendations based on novel knowledge that were not represented in the BCKM. For example, considering the patient profile presented in the two previous sections, the TDM1 drug was not initially included in the BCKM. As a first step, we had to add this concept to the ontology as a child of 'Single Agent Targeted Therapies'. After adding the concept, we edited the rule in the GL-DSS that recommended continuing with Trastuzumab therapy and added the condition that the response to neoadjuvant treatment must be complete. Finally, we added a new rule that recommends TDM1 treatment in the case of a non-complete response to neoadjuvant treatment. Figure 5.5 illustrates this process.

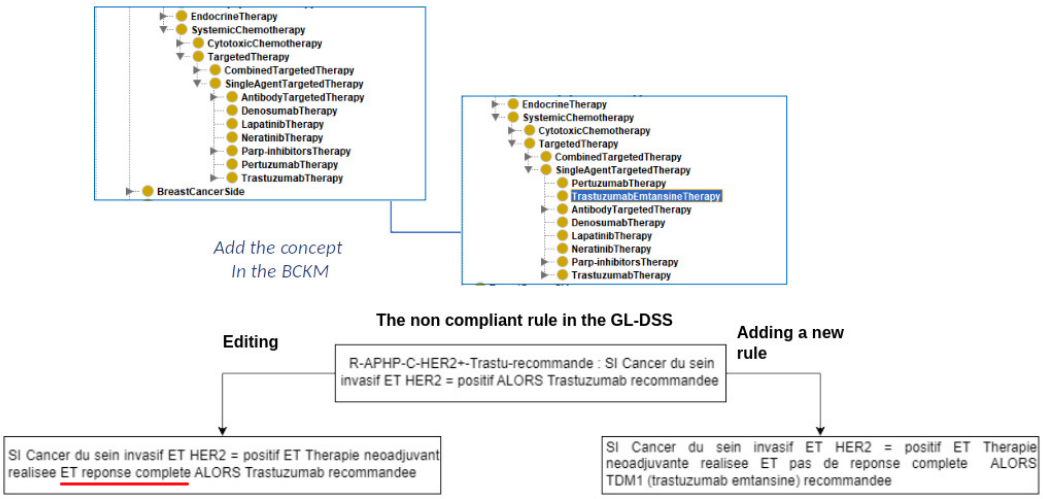


Figure 5.5: Adding concepts in the BCKM and rules in the GL-DSS knowledge base

5.2.3 Evaluation on complex cases

We also wanted to assess our hypothesis that the guidelines do not cover complex cases. So we tested the GL-DSS on a small corpus of complex cases and compared its performance to the performance obtained on non-complex clinical cases.

5.3 Results and discussion

The mapping process from the structured data obtained in chapter 3 and the BCKM ontology resulted in identifying a total of 71 attribute classes and 96 value classes. An example of the mapping for the BI-RADS classification is illustrated in figure 5.6. The complete mapping table can be found in Appendix C.

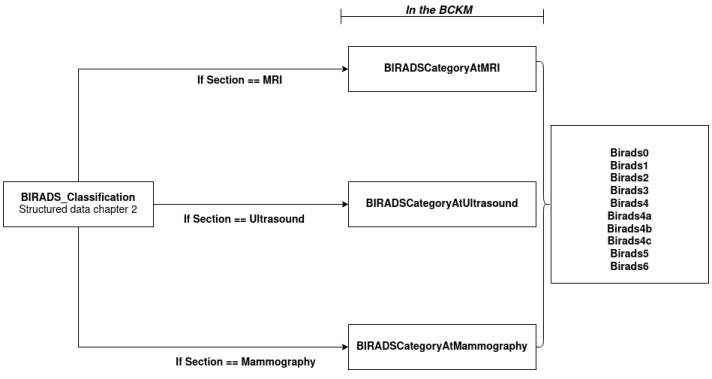


Figure 5.6: Exemple of mapping for the BI-RADS category

We selected a corpus of 160 BCPSs representing non-complex cases discussed during MTBs of Tenon Hospital in 2021 and covering all breast cancer scenarios (check section 3.2.4.3), 50 in

Scenario A, 35 in Scenario B, 25 in Scenario C and 50 from scenario D.

5.3.1 Comparison of MTB decisions and guidelines recommendations

5.3.1.1 Comparison of MTB decision to the GL-DSS recommendations

Among the 160 BCPS, we observed (Table 5.3.1.1) that MTB decisions were included among the GL-DSS recommendations for 121 cases (76%). For the other 39 cases (24%) there was at least one treatment step proposed by the MTB, that was not included in the GL-DSS recommendations.

Table 5.1: Comparison of $\mathbf{D}^{\mathbf{MTB}}$ and $\mathbf{R}^{\mathbf{GL\text{-}DSS}}$

	Scenario A	**Scenario B**	**Scenario C**	**Scenario D**	**Total**
Total	50	35	25	50	160
$\mathbf{D^{MTB} = R^{GL\text{-}DSS}}$	39	23	17	42	121
$\mathbf{D^{MTB} \neq R^{GL\text{-}DSS}}$	11	12	8	8	39
% of non-compliance	22%	34%	32%	16%	24%

We suggest that the cases where the GL-DSS recommendations followed MTB decisions (76%) represent clinical profiles for which the evidence remains unchanged, thus obviating the necessity of updating the knowledge base in this context.

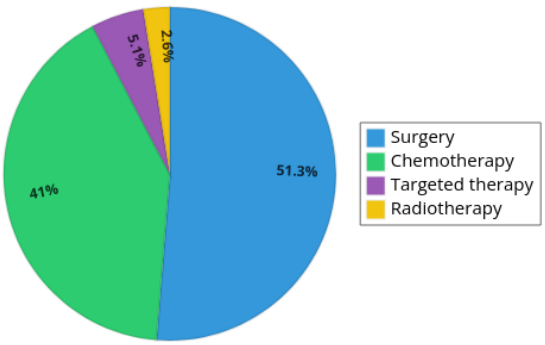


Figure 5.7: Distribution of the cases for which MTB decisions were not among the GL-DSS recommendations by modality of treatment.

Cases for which MTB decisions were not among the GL-DSS recommendations : Let us focus on the cases for which MTB decisions were not among the GL-DSS recommendations. As previously mentioned, we propose that instances where $D^{MTB} \neq R^{GL\text{-}DSS}$ may potentially indicate situations where there has been a progression or modification in the guidelines. As depicted in Figure 5.7, the majority of modalities observed in the profiles for which MTB decisions were different from GL-DSS recommendations were surgical interventions (51%) and chemotherapy (41%), there were no instances of difference concerning hormone therapy, while only one case exhibited difference regarding radiotherapy, and two cases exhibited difference with targeted therapies.

Table 5.2 describes the patient profiles for all these casesr and the differences between MTB decisions and GL-DSS recommendations. By conducting this comparison, we have obtained a comprehensive overview of potential shifts in practices when MTB decisions deviated from GL-DSS recommendations. To confirm that these deviations truly represent practice evolutions, we

can further examine MTB decisions for these cases and compare it to the recommendations provided by SENORIF.

5.3.1.2 Comparison of MTB decisions to the SENORIF recommendation

Of the 39 BCPSs where the recommendations of the GL-DSS did not comply with the MTB decision, SENORIF's recommendations were in line with MTB decisions for 38 cases (as we can see in table 5.2), and not in line for one case (one in which SENORIF did not recommend radiotherapy as decided by the RCP physicians). After discussing the clinical case with the experts, it was re-classified as a complex case.

Table 5.2: Results of the comparison by treatment modality

Patient Profile	$\mathbf{R}^{\text{GL-DSS}}$	$\mathbf{R}^{\text{SENORIF}}$
	Surgery (20 BCPSs)	
Recurrent invasive cancer	No indication for mastectomy	Indication for mastectomy + axillary lymph node dissection (ALND), no sentinel lymph node biopsy (SLNB)
Recurrent in situ cancer	No indication for mastectomy	Indication for mastectomy + SLNB
Multifocal cancer	Indication for mastectomy and ALND	No indication of mastectomy and no indication of ALND, possibility to do SLNB + Lumpectomy
Patients with BRCA mutation	No indication for mastectomy	Indication for mastectomy
	Chemotherapy (16 BCPSs)	
Triple-negative cancer with Tumor size ≥2cm and/or N of TNM >1	No indication for neoadjuvant chemotherapy ($\text{Chemo}^{\text{NeoAdj}}$)	Indication for $\text{Chemo}^{\text{NeoAdj}}$
Triple-negative cancer T1cN0	No indication for $\text{Chemo}^{\text{NeoAdj}}$	$\text{Chemo}^{\text{NeoAdj}}$ to be discussed based on clinical parameters (e.g., tumor-to-breast size ratio)
Negative HER2 and hormone receptor-positive	No specific indication based on genomic testing	Indication for Oncotype DX, a genomic test that evaluates the benefits of adjuvant chemotherapy ($\text{Chemo}^{\text{Adj}}$)
Triple-negative cancer after an incomplete response to $\text{Chemo}^{\text{NeoAdj}}$ and post-surgery	No indication or specification for $\text{Chemo}^{\text{Adj}}$	Indication for $\text{Chemo}^{\text{Adj}}$ with Capecitabine

	Targeted therapy (2 BCPSs)	
HER2-positive cancer with Tumor size ≥2cm and/or N of TNM >1	No indication for neoadjuvant therapy ($\text{Therapy}^{\text{NeoAdj}}$)	Indication for $\text{Therapy}^{\text{NeoAdj}}$
HER2-positive cancer after an incomplete response to $\text{Chemo}^{\text{NeoAdj}}$ and post-surgery	No indication or specification for adjuvant therapy ($\text{Therapy}^{\text{Adj}}$)	Indication for $\text{Therapy}^{\text{Adj}}$ with TDM1 (trastuzumab emtansine)
	Radiotherapy (1 BCPS)	
Elderly patient, SBR 2, pT1cN0, no emboli,Negative HER2 and hormone receptor-positive	Indication for radiotherapy	No indication for radiotherapy

5.3.2 Identification of updates in GL-DSS's knowledge base

By analyzing the 38 profiles where $\mathbf{D}^{\text{MTB}} \in \mathbf{R}^{\text{SENORIF}} \notin \mathbf{R}^{\text{GL-DSS}}$, we were able to identify specific elements that suggest modifications of existing rules. A comprehensive overview of these findings is presented in Table 5.3. For a more detailed account of all the identified rules, please refer to Appendix B (the document is in French, rules and guidelines are also in French).

Table 5.3: Results of knowledge base update

Modality	**Number of modifications**	**Number of additions**	**Number of deletions**	**Total**
Surgery	6	11	6	23
Chemotherapy	10	7	3	20
Targeted Therapy	1	3	0	4
Radiotherapy	1	0	0	1
Total	**18**	**21**	**9**	**48**

A total of 18 rules were modified, 9 rules were deleted and 21 new rules were added. Here are examples of each case:

1. **Update of existing rules:** For instance, in cases of multifocal tumors, the current GL-DSS's rule suggests ALND based on the 2016 AP-HP guidelines, as multifocality is considered a contraindication for SLNB. In contrast, SENORIF guidelines do not consider multifocal or multicentric cancers as a contraindication for the sentinel lymph node biopsy. Thus, the rule is modified as follows: *"R-2016APHP: If Invasive Breast Cancer AND Bifocal Lesion AND distance > 20mm THEN SLNB Contraindicated"* → *"R-SENORIF: If Invasive Breast Cancer AND Bifocal Lesion AND distance > 20mm THEN SLNB possible."*

2. **Addition of new rules:** For example, a rule is added concerning mastectomy in cases of recurrence: *"If InSitu carcinoma AND relapse THEN Mastectomy + SLNB."*

3. **Removal of outdated rules:** For instance, the rule *"If non-invasive breast cancer AND bifocal lesion AND distance > 20mm AND cN0 AND not cT4d THEN Recommended Mastectomy*

(p18)" is removed since, according to SENORIF, a multifocal cancer is not an indication for mastectomy.

5.3.3 Evaluation on complex cases:

To asses our hypothesis that the guideline-based CDSSs are not efficient in complex cases (section 1.3), we tested the *2016-APHP-based* GL-DSS on a corpus of 40 complex cases. Among these cases, 6 were not considered because the histologic type was not cancer (but they were still complex cases). Among the remaining 34 cases, the GL-DSS's recommendations were not compliant with MTB decisions in 23 cases (68%), which is considerably higher than the 23% non-compliance rate obtained with the same system on non-complex cases.

5.3.4 Discussion

As guidelines evolve, new versions often lack explicit indications of changes. We are far from the concept of "living guidelines" which emphasizes continuous updates and transparency in modifications (Li *et al.*, 2022). Consequently, the absence of a structured mechanism to identify and track changes in the guidelines makes it challenging to efficiently update a CDSS's knowledge base and keep it aligned with the most current evidence and recommendations.
This lack of consistency can lead to significant variations in format, structure, and terminology between different guideline versions, posing challenges for comparisons using natural language processing methods. The diverse recommendations and approaches across guidelines may result in discrepancies, requiring thorough analysis and a deep understanding of specific contexts. Moreover, the regular updates of guidelines add a temporal dimension to the task, demanding continuous monitoring of new versions and their respective modifications.

The proposed method involves using real-world data to detect evolutions in medical practices. By comparing the decisions of the multidisciplinary teams to the outputs of the CDSS, we were able to identify profiles where there have potentially been changes in medical practices. This comparison process efficiently captured the parts of the SENORIF guidelines where these evolutions of practices are described. Through this method, decision rules that needed to be added, modified, or removed from the GL-DSS knowledge base were concretely identified. Additionally, new concepts related to emerging treatments or decision variables were incorporated into the BCKM ontology. By leveraging real-world data and MTB decisions, this approach was effective in detecting and adapting to changes in medical practices as reflected in the latest guidelines. However, even if the comparison is easily automated, we still need to manually check the SENORIF guidelines to update the knowledge base.
In fact, during the manual analysis of SENORIF, we discovered practice evolutions that had not been expressed in the set of clinical profiles we examined. This emphasizes the importance of complementing this approach with additional mechanisms for monitoring and regularly updating new versions of the guidelines. Thus, to ensure more comprehensive coverage of medical practice evolutions, it is essential to combine the comparative method with careful scrutiny of new guideline versions and continuous monitoring of changes.

One of the primary challenges faced in this study was the quality of the data obtained from the clinical notes. In addition to the issues outlined in Chapter 3 regarding the structure and writing style of the BCPS, we encountered further difficulties when manually creating patient profiles in the BCKM format.
One particular challenge arose from the fact that some attributes required by the decision rules in

the GL-DSS were not consistently present in the textual BCPS. These missing attributes included TNM classification, SBR grade, the distance between tumors in the case of multicentric tumors, and tumor size before and after neoadjuvant treatment, among others. As a result, we had to invest significant time in understanding patient profiles, and at times, we were compelled to complete missing information based on the text to ensure the proper functioning of the system. For example, a frequent occurrence was the TNM status, which was not always expressed as N0 or N1 but provided in a textual form, such as "no axillary invasion.".
Moreover, we also observed numerous errors in the BCPS, such as the reversal of right and left sides, tumor size expressed in meters instead of centimeters, and other inconsistencies. Addressing these errors further added to the time-consuming nature of the manual data processing process. These challenges highlight the fact that even if we have an NLP pipeline that works well in extracting valuable information, work has to be done to ensure top-quality data during the BCPS creation process

Finally, the utilization of the GL-DSS in DESIREE for non-complex breast cancer cases has yielded highly promising results. In fact, despite being implemented on guidelines from 2016, in only 24% of the cases (39 out of 160), the recommendations generated by the GL-DSS differed from MTB decisions. However, it is important to note that there were 10 cases where the MTB decided a mastectomy because there was a relapse. Considering that the GL-DSS was made for the management of primary breast cancer care, there were no rules that deal with patients having a relapse, even if in the AP-HP 2016 guidelines, mastectomy was recommended for patients with relapse. So these specific cases do not represent an evolution of the guidelines, but, we considered these rules regarding patients with relapse when updating the knowledge base. for the remaining 29 cases where there was a difference between MTB decisions and GL-DSS's recommendations, the divergence is primarily attributed to practice evolutions, which can be easily addressed through the addition or modification of rules in the knowledge base. The only instance where the GBP recommendations did not align with MTB decisions was for a patient classified as a non-complex case but was actually a complex case (a transgender patient with a complex hormonal treatment decision). Therefore, even though this requires confirmation by updating the GL-DSS knowledge base and using it on a larger sample, the system's performance is really promising for non-complex cases.

It is important to mention that this approach does have limitations. If guidelines propose new management approaches for cases previously considered compliant, MTB decisions may potentially become non-compliant with the most recent guidelines. This issue could be more important if guidelines evolve at a faster pace than the GL-DSS. As a potential solution, integrating feedback mechanisms into the GL-DSS could be beneficial in identifying cases where compliant decisions become non-compliant, thus allowing for timely updates to maintain alignment with current guidelines.

5.4 Conclusion

In conclusion, regularly updating a knowledge base requires an in-depth analysis of clinical practice guidelines. Despite the encountered challenges, the semi-automated approach employed in this study allowed for the concrete observation of medical practice evolutions and the identification of necessary modifications in the knowledge base. However, issues regarding data quality and the need for further automation underscore the importance of establishing standards for presenting clinical information in BCPSs.

Furthermore, the utilization of the GL-DSS on non-complex breast cancer cases has shown highly promising results. Conversely, its performance on complex cases demonstrated poorer outcomes. This finding supports our hypothesis that distinguishing between non-complex and complex cases could ensure a more robust response from the guideline-based system, thereby enhancing the potential for routine clinical use. The ongoing development of such knowledge-based systems holds significant potential in supporting healthcare professionals during the decision-making process and ultimately improving patient care, provided the knowledge base remains up-to-date.

CHAPTER 6

Case-based decision support

Clinical practice guidelines generally do not offer appropriate guidance for managing complex patient cases. As a result, guideline-based decision support often falls short, prompting clinicians to seek alternative approaches such as patient similarity-based decision support. This research compares two methods for calculating the similarity between breast cancer patients. The first method employs one specific type of measure to calculate the similarity between two cases based on their attributes (termed "single-measure" method), while the second method, termed the "hybrid method", utilizes distinct similarity measures tailored to different attribute categories. Expert knowledge is incorporated through a weighted average of attribute measures, enhancing both the interpretability and performance of case similarity determination. Optimization techniques were employed for both methods, utilizing deep metric learning for the "single-measure" approach and classical (non-machine learning) optimization methods for the hybrid method. The dataset employed for this work was a cohort of 100 arbitrarily chosen BCPSs for training purposes, and the method was evaluated on 10 randomly selected complex BCPSs.

Although deep metric learning methods were explored to optimize the "single-measure" approach, the results did not meet initial expectations, indicating the need for further refinement. Our research underscores the importance of selecting appropriate similarity measures according to the nature of the attributes and of effectively weighting them based on the type of variables used. This approach substantially improves the accuracy of patient similarity assessments and facilitates the comprehensive interpretation of the results by integrating expert knowledge.

6.1 Introduction

Case-based reasoning (CBR) is a problem-solving methodology that draws on past experiences (cases) to solve new problems (De Mantaras, 2001). In the context of healthcare, CBR has emerged as a valuable approach for clinical decision-making, providing a framework to guide current diagnostic and treatment decisions. CBR offers healthcare professionals a systematic method to decision-making by leveraging the knowledge and experiences captured in the case base. It promotes personalized medicine by taking into account patient-specific characteristics and facilitates the development of targeted and effective treatment strategies. Furthermore, the evaluation of patient similarity and the advancement of similarity measures play a major role in enabling precision medicine and supporting data-driven decision-making in real-world healthcare settings.

In the management of complex cases, as mentioned in section 1.1.3, clinicians desired retrieving similar patient profiles as the decisions made for these similar profiles might be similar to the decisions to be made for the complex case. Therefore, we suggested that the implementation of CBR can offer significant advantages. Complex cases often present unique challenges, requiring a more personalized and tailored approach to decision-making. By utilizing CBR, healthcare professionals can tap into a vast repository of past cases with similar clinical profiles or characteristics. This allows for the retrieval of relevant cases that offer valuable insights and potential solutions specific to the complexity of the current case.

As seen in section 2.3, CBR reasoning process involves several steps. Firstly, relevant cases are **retrieved** from a case repository based on their similarity to the current patient case. Similarity measures are employed to identify cases with similar clinical profiles. Once relevant cases are retrieved, the next step is the **reuse** of the retrieved information. Decisions made in similar cases are adapted and applied to the current patient case. This step allows healthcare professionals to capitalize on previous successful interventions or treatments, potentially saving time and improving the quality of care. The third step in CBR is the **revision** of decision. The retrieved decision is examined and adjusted to fit the specific context of the current patient case. This step ensures that the decision is tailored and optimized for the individual patient's needs, considering factors such as comorbidities, preferences, and available resources. Lastly, the revised decision, if validated as a new decision for the patient, is **retained** in the case base for future reference. This step contributes to accumulating experience over time, as the case base grows with additional cases and their corresponding decisions.

In this chapter, our focus is on the development of a method for recalling patients similar to a given patient to propose therapeutic options specifically tailored to her needs. With the aim to enhance the decision-making process by leveraging CBR techniques, we compared two methods. The first one, the "single-measure method", is based on using one specific measure (e.g., Cosine, Euclidean distance) for all attributes to calculate the similarity between two patients. The second one, called the "hybrid method", uses a weighted average of multiple measures to assess patient similarity. Additionally, we used optimization techniques to enhance the precision and accuracy of the similarity calculation algorithms, evaluated on an expert-curated patient case dataset.

Through this chapter, we aim to contribute to the field of decision support in healthcare by presenting a comprehensive methodology that takes advantage of CBR techniques for recalling similar breast cancer patients and proposing personalized therapeutic options to any new complex breast cancer patient.

6.2 Methods

The methodology is centered around the use of a corpus of textual BCPSs that have undergone transformation into coded data using the structured data extraction algorithm described in Chapter 3. The structured data model utilized here aligns with the one previously described in Section 3.2.1.2. Figure 6.1 shows an overview of the pipeline implemented in this chapter :

- We began by constructing two datasets, the *similarity learning dataset*, consisting of 100 randomly selected BCPSs (stratified as 50 complex cases and 50 non-complex cases) that were grouped into clusters by an expert and used for training the algorithms. Then another subdataset made of 10 randomly selected complex BCPSs was used to evaluate the algorithm, an expert calculated for each of the 10 selected patients, the top 5 similar patients from the *similarity learning dataset*. The resulting dataset was used as a gold standard for evaluating the similarity measure and named *similarity evaluation dataset* (Please refer to figure 1.2. for a graphical visualization)
- Next, we proceeded to create a generic similarity measure to assess the similarity between patients. We compared the performance of both the single-measure and the hybrid method to calculate similarity.
- Then, we optimized the similarity calculation for both methods. We explored traditional optimization techniques and the use of deep learning.
- Finally, both methods were evaluated on the constructed gold standard.

6.2.1 Dataset building

We collaborated with an advanced oncology expert to develop a dataset for a similar case detection task. To start, we incorporated the concept of scenarios (described in section 3.2.4.3) into this methodology. Based on discussions with the expert, we concluded that patients can only be compared and considered similar if they are in the same scenario. For example, a patient who has already undergone surgery (scenario D) cannot be compared to a patient who is in the initial decision stage (scenario A). Taking this into consideration, we made sure the dataset included only patients in scenario D. Indeed, we found that Scenario D was the most common and encompassed a wide range of attributes. So in this chapter, all the BCPSs described represent patients who have already undergone surgery without a neoadjuvant treatment (scenario D).

A corpus of 100 BCPSs in scenario D was selected among the 1,048 BCPSs from the *complexity learning dataset* (refer to figure 1.2), and stratified into two groups: 50 BCPSs of non-complex cases and 50 BCPSs of complex cases. The advanced expert manually classified the 100 BCPS into clusters of similar patient situations based on their clinical characteristics but not on the decisions that were made for these patients (the clusters here are not complex or not complex but represent clusters of patients with shared key clinical characteristics). This clustered dataset, named *similarity learning dataset*, served as the training set for similarity learning.

The sample of 10 complex BCPSs was randomly selected from the same source (*complexity learning dataset*) and chosen to be different from the 100 BCPSs of the *similarity learning dataset*. Following this, for each of the chosen 10 complex BCPSs, the expert thoroughly analyzed the clinical context, then identified the 5 cases that exhibited the highest resemblance to the selected BCPS from the *similarity learning dataset*. This resulting subset was utilized as a gold standard for evaluation and is named *similarity evaluation dataset*.

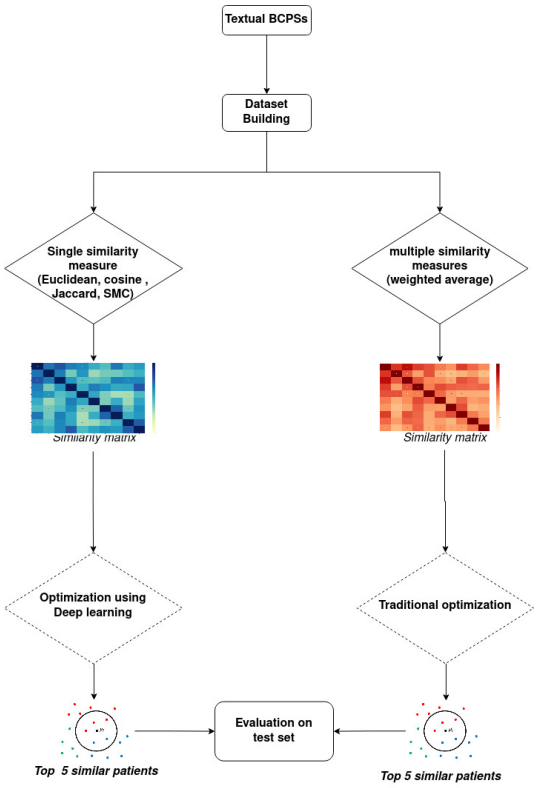


Figure 6.1: Pipeline used for CBR decision support

Refining Attribute Values and Establishing Ordinal Variables During this phase of dataset development, we worked with the expert to refine some attribute values and transform them into meaningful ordinal variables. The objective was to improve the quality of the dataset and facilitate the similarity and dissimilarity calculation between 2 patients, by ensuring that attribute values represented significant differences.

To illustrate this process, let's consider the T attribute of the TNM classification, which describes the tumor.
We establish a classification scheme for the T attribute. The scheme assigns numerical values to different T classifications, reflecting the severity or progression of the tumor size. For instance, we categorized the values as follows:

1. (Tx): Tumor size cannot be assessed.
2. (T0): No evidence of a primary tumor.
3. (Tis, T1mic): Carcinoma in situ or microinvasion.
4. (T1abc): T1 tumors (<= 2 cm).
5. (T2, T3): T2 or T3 tumors (> 2cm).
6. (T4): Tumor of any size with direct extension to the chest wall or skin.

By organizing the attribute values in this manner, we effectively captured the increasing severity of the tumor as the numerical values progressed.

Furthermore, we applied a similar methodology to other attributes within Scenario D, including the N (Node) attribute. By refining the attribute values and transforming them into ordinal variables, we ensured that the dataset represented the clinical reality and facilitated precise clustering of breast cancer patients. In addition to that, we also grouped the patients according to the nature of the tumor, indeed we distinguished between patients with invasive breast cancer and patients with InSitu breast cancer.

In the end, we had a training dataset of 100 scenario-D BCPSs represented in a CSV table containing the needed attributes for this scenario, and the cluster assigned to each BCPS by the expert. In addition to that, we had a test set of 10 BCPSs representing complex cases from the same scenario. For each of these 10 BCPSs, rather than having the cluster labeled for each patient, we had the 5 most similar BCPSs (if any) within the training dataset.

6.2.2 Construction of generic similarity measures

Following a literature review and considering the type of variables in the data model, we have chosen to use the similarity measures listed in Table 6.1.

Table 6.1: Selected similarity measures: *A brief description of the variables and formulas used. depth(s1) and depth(s2) represent the distance from the root to the nodes s1 and s2, respectively. Depth (lsc(s1, s2)) represents the distance from the root to the common branch of concepts s1 and s2.*

Variable	**Measure**	**Formula**
Numeric (e.g Tumor size)	Cosine Similarity	$\frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$
Categorical (e.g Birads score)	Jaccard Similarity	$\frac{A \cap B}{A \cup B}$
Ordinal (e.g Tumor grade) or numeric	Euclidean Similarity	$\frac{1}{1+\sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}}$
Hierarchical (e.g Histologic type)	Wu and Palmer Similarity	$SimWP(s_1, s_2) = 2 \times \left(\frac{depth(lcs(s1,s2))}{depth(s1)+depth(s2)}\right)$

To calculate the similarity between cases, we have used two methods (see Figure 6.1 for a visual comprehension of the pipeline):

- Method 1: Utilizing a unique measure for the calculation: "single-measure" method.
- Method 2: Utilizing multiple measures by taking a weighted average: "hybrid" method.

6.2.2.1 Method 1: Utilizing a unique measure for the calculation

In this method, a single similarity measure is applied to all attributes without variable grouping, with a normalization step:

- For Cosine and Euclidean measures, we normalized all ordinal variables on a scale of 0-1 to ensure they had the same scale as binary variables. The categorical variable, histological type, was customized to have two binary variables: invasive carcinoma (yes/no) and carcinoma in situ (yes/no).

- For the Jaccard measure (a measure for categorical data), the numeric attribute age was transformed into a binary variable: elderly (yes/no), setting the threshold at 75. We removed the numeric variable for tumor size and only kept it with the categorical variable of T from the TNM classification.

For this method, the input for each similarity measure was a vector representing the values for each attribute according to the information extracted from the BCPS. It gives as output a similarity matrix containing the values of similarity between each pair of patients.

6.2.2.2 Method 2: Utilizing multiple measures by taking a weighted average

A. Variable grouping In this method, instead of normalizing the attributes to fit with a single similarity measure, we grouped the variables according to their type and we used the similarity measure that fits with that type, i.e. Jaccard for the categorical variables (including binary variables), Euclidean for the ordinal variables, and Wu & Palmer for hierarchical variables. Table 6.2 below describes which measure is utilized for the different attributes mentioned in 6.3.

Table 6.2: Selected similarity measures and corresponding attributes

Variable	Measure	Variables
Categorical	Jaccard	Vulnerable, Hormone Receptors, HER2, Triple Negative, Comorbidities, Bilateral Cancer, Recurrence, Menopausal Status, Carcinoma in Situ Associated, Focality, Margins, Presence of Other Cancers, History of Radiotherapy, Genetic Mutation, Type of Surgery, OncotypeDX Value
Ordinal and Integers	Euclidean (Normalized)	Ki67, T, N, Grade of Invasive Tumor, Grade of In Situ Tumor, Age, Size of tumor
Hierarchical	Wu and Palmer	Histological Type (represented in figure 6.2)

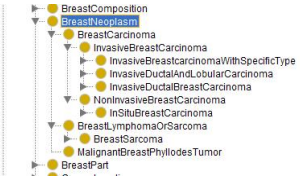


Figure 6.2: Representation of histologic types in the BCKM ontology

Implementation of Wu & Palmer similarity : In order to implement the Wu and Palmer method using the ontology, we used a programming module for ontology-oriented programming in Python 3 called OwlReady2 (Lamy, 2017) . This module allowed to import the BCKM ontology into Python, enabling to work with the hierarchical structure of the possible values of the "histological type" variable. The Wu and Palmer similarity measure provides a formula (see Table 6.1) for calculating a similarity between two concepts in a hierarchy by considering the depth of each class and the depth of their common parent class.

B. Hybrid similarity calculation function After grouping the attributes according to the similarity measure, we implemented a Python function that takes for each BCPS the vector of attribute values described in table 6.3. The function takes weights for each attribute, allowing for the variation of weights based on the importance assigned to the attribute. Finally, a similarity measure between 0 and 1 is calculated for each pair of patients by taking a weighted average of the measures

across all attributes. This function facilitates the computation of similarity scores and enables the comparison of cases based on a comprehensive set of attributes.

C. Similarity matrix and hierarchical clustering Once the similarity function was implemented, we applied it to the dataset to obtain a similarity matrix for each pair of patients. The similarity matrix was then processed using the *fcluster* function, which performs hierarchical clustering and assigns cluster labels to the data points based on the desired number of clusters. We set the desired number of clusters to the number of clusters identified by the expert (18 clusters on the *similarity learning dataset)*. Clusters were computed from the matrix using the "ward" linkage method (Großwendt *et al.*, 2019). This method merges similar observations while minimizing the loss of inertia, promoting the formation of clusters of similar sizes. Results are visualized using a dendrogram.

After obtaining the dendrogram, we varied the weights of the similarity function to calculate the similarity matrix. The objective was to determine if it was possible to replicate the clusters identified by the expert by manually adjusting the weights assigned to each variable in the similarity function. Note that we chose to vary the weights only for the "hybrid method" because when comparing the results of the "hybrid method" without weights (all weights equal 1) to the "single-measure" method, we saw that the hybrid method gave better results.

To manually determine the variable weights that aligned with the gold standard results, we employed an interactive visualization framework such as "*Bokeh.*" This framework facilitates the visualization of the results (see figure 6.3). The iterative process of adjusting the weights aimed to identify the combination that yielded clusters similar to those annotated by the expert, thereby refining the similarity calculation and improving the accuracy of patient clustering.

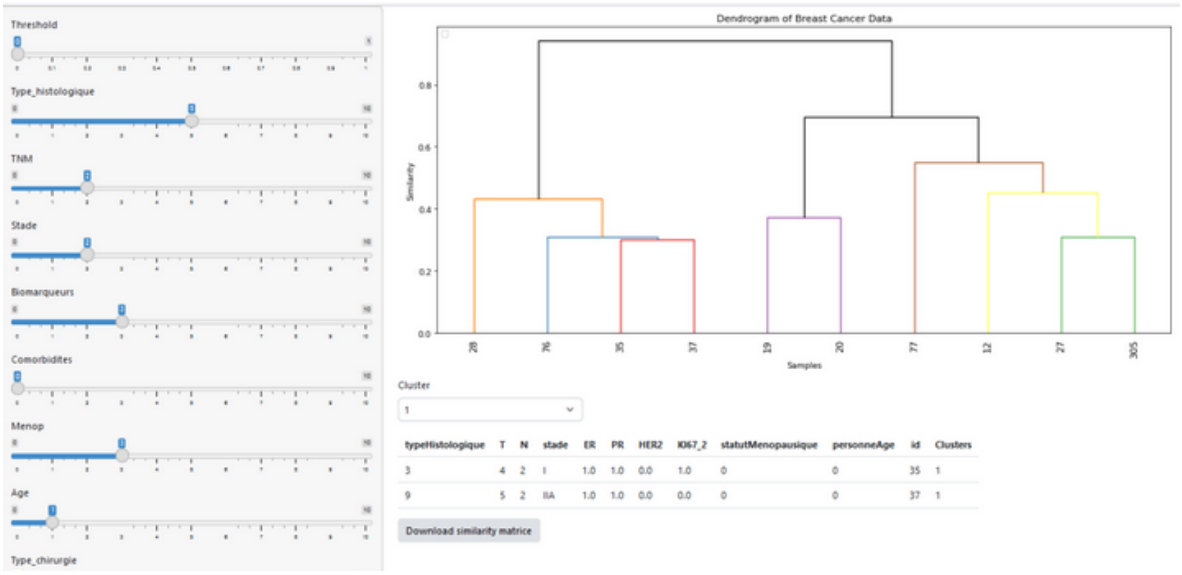


Figure 6.3: Dashboard for manual weights definition

6.2.3 Similarity learning

After obtaining the similarity matrix using different measures, we aimed to optimize the weights and use learning methods to improve the clustering results. While manually searching for the

weights that closely match the expert clustering can be helpful, it is essential to find an automatic method for this learning process.

To address this optimization task, we used the *similarity learning dataset* (100 curated and structured BCPSs) and we considered the generic similarity measures obtained using both Method 1 (section 6.2.2.1) and Method 2 (section 6.2.2.2):

For Method 1, where only one similarity measure is used, we employed deep learning methods to learn suitable embeddings for the initial patient attributes. Specifically, we utilized Siamese networks (Tran *et al.*, 2020), which are neural network architectures commonly used for similarity calculation tasks. Siamese networks learn to project the patient attributes into a shared embedding space, where the similarity between patients can be measured based on the distance or similarity between their embeddings. By training the Siamese network using appropriate loss functions, we can optimize the weights to achieve better similarity assessment and clustering results.

For Method 2, the similarity function described in Section 6.2.2.2.3 utilizes specific similarity measures based on the variable type. Consequently, utilizing a neural network-based method for optimization in this case becomes impractical. Neural networks modify the initial embeddings through their backpropagation logic, resulting in completely different embeddings from the initial vectors for each patient. Therefore, using the custom similarity function becomes unfeasible. Thus, for Method 2, we employed optimization techniques that do not rely on neural networks. We used the Optuna library (Akiba *et al.*, 2019) to randomly sample weight values for the similarity function and evaluate their performance on the clustering task. The weight values that yielded the best clustering results were used to define the similarity function.

6.2.3.1 Similarity learning for method 1 (the "single-measure" method)

For Method 1, we employed the following steps for embedding learning with Siamese networks using triplet loss as a loss function. Figure 6.4 resumes the algorithm described below.

Siamese network architecture We designed a Siamese network architecture with Multi-Layer Perceptrons (MLPs). Siamese networks are effective in learning similarity relationships by projecting patient attributes into a shared embedding space. This shared embedding space enables meaningful comparisons between cases based on the distance or similarity between their embeddings.

Training initialization To initiate the training process, we selected a small sample for the *similarity learning dataset* of 18 cases, ensuring representation from each cluster (18 clusters). These patients served as an initial sample to start the training.

Iterative training We performed iterative training by adding one patient at a time to the training process. For each patient, the following steps were executed:

1. Extraction of Triplets: Each time we added a patient, we calculated all possible triplets in the current batch. This was done using the batch-all strategy (Li *et al.*, 2021), where we considered all combinations of anchor-positive-negative triplets. As explained in section 2.3.2.1 valid triplets are triplets where the anchor and the positive sample are in the same cluster, while the anchor and negative sample are in different clusters.

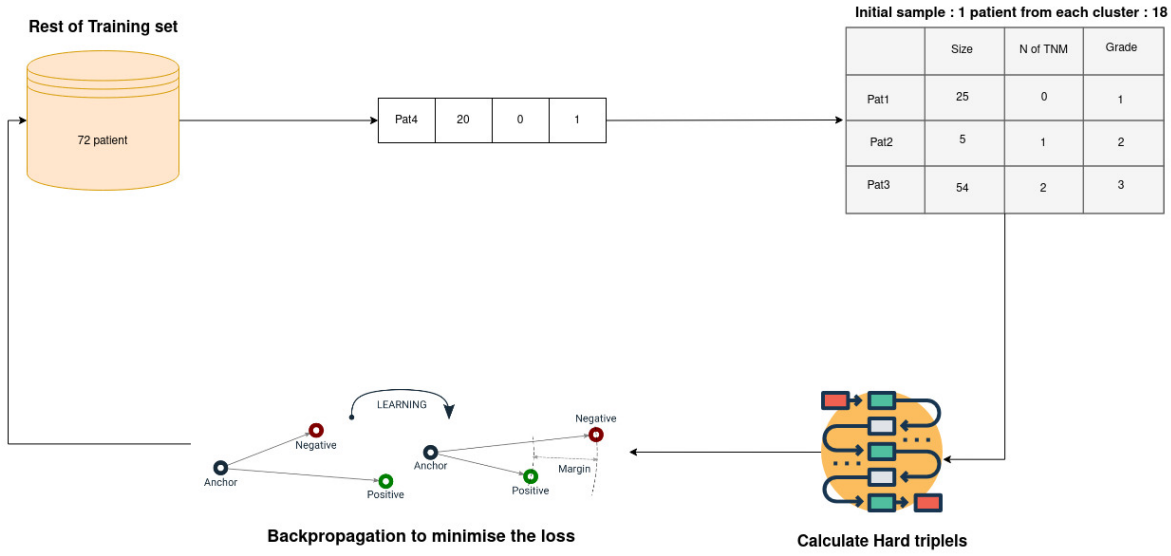


Figure 6.4: Pipeline for similarity learning using neural networks

2. Calculation of triplet loss: The triplet loss was computed for each triplet in the batch using the Siamese network and the following loss function:

$$L = \max(d(a,p) - d(a,n) + \text{margin}, 0)$$

where $d(a,p)$ represents the distance between the anchor and positive embeddings, $d(a,n)$ represents the distance between the anchor and negative embeddings, and margin is a predefined margin value. The triplet loss encourages the embeddings of anchor-positive pairs to be closer than the embeddings of anchor-negative pairs by at least the margin value.

3. Update of the Siamese Network: We updated the weights of the Siamese network using backpropagation and the computed triplet loss to improve the embeddings. The objective was to minimize the average triplet loss across all the triplets in the batch.

This iterative training process was repeated for each additional patient in the dataset. Our aim was to learn embeddings that accurately represented patient attributes and captured the desired similarity relationships. To measure similarity between two embeddings, we utilized Euclidean similarity, a widely-used metric in Siamese network architectures and the most used one for CBR systems in healthcare according to a recent review (Noll *et al.*, 2022).

6.2.3.2 Weights optimization for method 2 (the hybrid method)

To automate the process of weight calculation in the similarity function of method 2. We employed the **Optuna** algorithm (Akiba *et al.*, 2019) , which is an optimization library that automatically finds the best parameters to maximize or minimize a given objective function.

In this context, the objective function is used to evaluate the performance of the similarity function based on the adjusted rand index (**ARI**) (Sundqvist *et al.*, 2020), it is a measure that ranges from -0.5 to 1.0, where a score of 1.0 indicates a perfect match or identical clusterings (up to a permutation). A value close to 0.0 indicates random labeling, independent of the number of clusters

and samples. In cases where the clusterings are particularly discordant or dissimilar, the ARI can have a value below -0.5, indicating a significant disagreement between the two sets of clusters. By iteratively optimizing the weights to maximize the ARI, we aim to enhance the clustering results and achieve better agreement with the expert annotations. The optimization process follows the steps outlined below:

1. Three weights are fixed based on the expert knowledge, with a high weight of 20 assigned to the variables "triple negative," "recurrence," and "geriatric frailty."
2. Optuna proposes integer weights ranging from 0 to 20 for the remaining variables.
3. A new similarity matrix is calculated using the weights provided by Optuna and the similarity function described in Method 2.
4. Clusters are calculated from the similarity matrix as explained in section 6.2.2.2.4
5. The resulting clusters were then evaluated using the ARI to measure the agreement between the expert-assigned clusters and the algorithm-generated clusters.
6. The process is repeated with the objective of maximizing the ARI.

The algorithm 1 below describes the objective function :

Algorithm 1 Objective Function for Weight Optimization

Require: $Data$: Training Dataset of 100 patients annotated by expert
Require: $encodedCategories$: Labels (clusters) of patients
  function objective($trial$)
    $weights \leftarrow []$
  for i **in** variables **do**
    if $i = TripleNeg$ **or** $i = Recurrence$ **or** $i = GeriatricFrailty$ **then**
      $weights$.append(20)
    else
      $weights$.append($SuggestInteger(1,20)$)
    end if
  end for
  $NewMatrix \leftarrow calculateSimilarityMatrix(Data, weights)$
  $newGroups \leftarrow fcluster(NewMatrix)$
  $ari \leftarrow AdjustedRandScore(encodedCategories, newGroups)$
  return ari
  end function

6.2.4 Evaluation

Once we obtained the similarity matrix from each method, we got 7 similarity matrices:

- $M1^{Euclidean}$, $M1^{Cosine}$, $M1^{Jaccard}$: represents the similarity matrices obtained using, respectively the Euclidean, Cosine, and Jaccard similarity as a unique measure for calculation, following the attribute values transformation step described in section 6.2.2.1.

- $M1^{\text{Siamese network}}$: The similarity matrix is obtained following the method described in section 6.2.3.1.
- $M2^{\text{NoWeights}}$: The similarity matrix is obtained by calculating the weighted average of the three measures following Method 2 (described in section 6.2.2.2), with a weight of 1.
- $M2^{\text{Weighted manually}}$: The similarity matrix is obtained following Method 2, with weights obtained manually.
- $M2^{\text{Weighted automatically}}$: The similarity matrix is obtained following Method 2, with weights obtained using the Optuna optimization.

We utilized the *similarity evaluation* dataset to evaluate the performance of the implemented methods. During the evaluation, we compared the top 5 patients annotated as similar by the expert (the gold standard) with the top 5 patients provided by the similarity methods. The evaluation metrics used for this comparison were precision and recall.

In the context of considering only the top 5 patients, precision and recall have the same value. The number of true positives represents the patients correctly identified as similar by the expert and the similarity methods, while the number of false positives indicates the patients falsely identified as similar by the methods. Similarly, the number of false negatives represents the patients who are actually similar but not included in the top 5 by the methods.

6.3 Results and discussion

6.3.1 Created datasets

As a result, we obtained 2 datasets of BCPSs representing patients in scenario D (after first surgery). The BCPSs were curated to keep the most important attributes. The list of attributes we took into consideration for scenario D and their associated values are specified in Table 6.3.

Table 6.3: List of attributes and their values for patients in scenarioD (surgery without neoadjuvant treatment)

Variables	Values
Age	Integer
Tumor size	Integer
Vulnerable	0 (no) / 1 (yes)
	0 (no) / 1 (yes)
Hormone Receptors	(Progesterone receptors, Estrogen receptors)
	0 (negative) / 1 (positive)
HER2	0 (negative) = score 0, 1, 2 and FISH negative
	1 (unknown) = score 2 and unknown FISH
	2 (positive) = score 2+, 3+, 2 and FISH positive
Triple Negative	0 (negative) / 1 (positive)
Ki67	0 (low) / 1 (high)
Comorbidities	0 (no) / 1 (yes)
Bilateral Cancer	0 (no) / 1 (yes)
Recurrence	0 (no) / 1 (yes)

Menopausal Status	0 (premenopausal) / 1 (postmenopausal)
Associated In Situ Carcinoma	0 (no) / 1 (yes)
Focality	0 (multifocal) / 1 (unifocal)
Clear surgical margins	0 (no) / 1 (yes)
Positive in situ margins	0 (no) / 1 (yes)
Positive invasive margins	0 (no) / 1 (yes)
Presence of Other Cancers	0 (no) / 1 (yes)
Antecedant of Radiotherapy	0 (no) / 1 (yes)
Genetic Mutation	0 (no) / 1 (yes)
Surgery Type	0 (Conservative surgery) / 1 (mastectomy)
Tumor Stage	1 (Tx) / 2 (T0) / 3 (Tis, T1mic) / 4 (T1abc) / 5 (T2, T3) / 6 (T4)
Lymph Node Status	1 (Nx) / 2 (N0, N0i-) / 3 (N0i+, N1mi) / 4 (N1, N2) / 5 (N3)
Invasive Tumor Grade	0 (unknown) / G1 / G2 / G3
In Situ Tumor Grade	0 (unknown) / G1 / G2 / G3
OncotypeDX Value	0 (not done/not yet) / 1 (RS<11) / 2 (11-25) / 3 (>25)
Histological Type	Hierarchical tree (see figure 6.2)

Insitu breast cancer As explained in 6.2.1. We distinguished between InSitu cancers and invasive cancers. Within the training set, there were 11 patients with InSitu breast cancer, who were also classified by the expert as InSitu.

6.3.2 Evaluation

Figure 6.5 shows the performance of all the methods implemented on the test set. The metric expressed is precision.

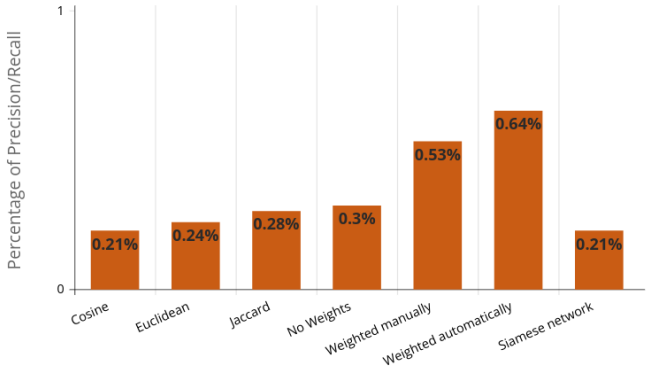


Figure 6.5: Precision of the different methods for CBR retrieval results

Regarding the single-measure method, the precision/recall scores hovered around 20% for Euclidean and Cosine to almost 30% for Jaccard, indicating the need for improvement (see figure 6.5).

In Method 2 - we wanted to assess if a hybrid approach with No Weights and Manual Weights improves the results. The hybrid method displayed some progress, showing slightly better results than the metrics in method 1, with a precision/recall score of 30% with no weights. And achieving 53% when calculating the weights manually.

Seeking to maximize each method's potential, we delved into optimization. While the Siamese Network (Method 1 optimization) did not show improvement using the Euclidean distance, employing the Optuna optimization algorithm for Method 2 achieved better results. It elevated the precision/recall score to approximately 64%, surpassing other methods.

6.3.2.1 Comparison of the different measures

Figure 6.6 shows the distribution of the extracted similar patients for each patient from the test set (there are 9 patients in the figure as there was a patient for which there were no similar cases), When looking at the results, we gained valuable insights into their effectiveness in identifying similar patients. Let's delve into the key observations:

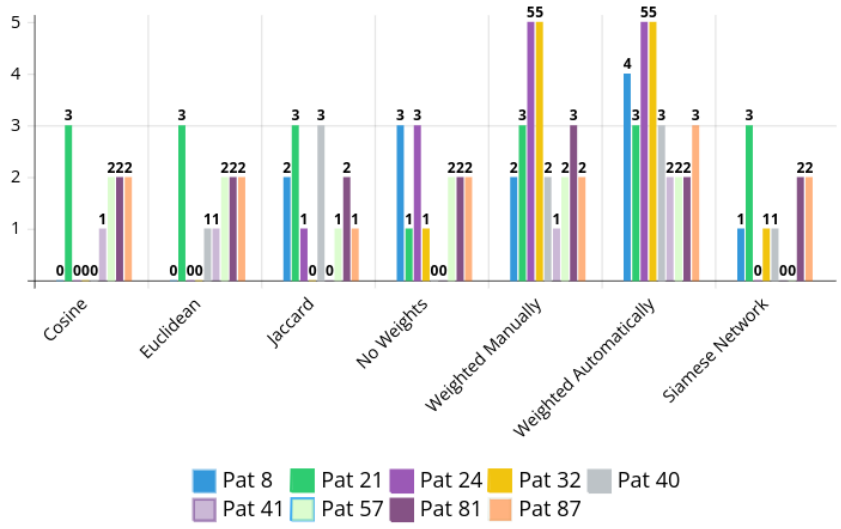


Figure 6.6: Distribution of top 5 extracted patients on the test set

- **Cosine and Euclidean Similarity:** These two methods exhibited strikingly similar results, with both yielding relatively low similarity scores for most patients. The scores mostly ranged from 0/5 to 1/5, indicating that these methods struggled to accurately identify similar patients in the dataset. For example, 4 patients (IDs Pat 8, Pat 24, Pat 40 and Pat 32) received scores of 0/5 using the Cosine metric, 3 of them also had 0/5 for Euclidean. Suggesting that the methods did not retrieve any of the similar patients classified by the gold standard.

- **Jaccard Similarity:** The Jaccard method performed slightly better than Cosine and Euclidean. It displayed improved scores for several patients, with 4 out of 9 patients receiving a similarity score of at least 2/5. However, it still fell short in effectively identifying similar patients across the dataset with 2 patients having a score of 0/5.

- **Weighted Average without Optimization (no weights):** Creating a hybrid metric that combines the similarity measures using a weighted average showed slight improvement over individual measures. Five out of the 10 patients in the test set had at least a score of 2/5.

- **Manually Weighted Method:** The manually weighted approach, where we carefully assigned weights to each similarity measure, showcased significant performance improvement. It

achieved high scores for a considerable number of patients. Notably, patients Pat 24 and Pat 32 received perfect similarity ratings of 5/5, highlighting their strong resemblance to other patients.

- **Automatically Weighted Method:** By leveraging the Optuna optimization algorithm to find optimal weights, and giving him the best manual weights as input to start the optimization, the automatically weighted method demonstrated further enhancement in similarity assessment. Reaching the best performance among all other methods.

- **Siamese Network:** The Siamese network method presented a mixed picture. While it exhibited improved scores for several patients, reaching up to 3/5, it struggled in accurately identifying similarities for others where the unique measure method did an initial good identification. For example, patient Pat 41 received a score of 0/5, while it had a score of 1/5 in the initial Euclidean similarity calculation. The network's architecture and training process might influence its performance, indicating the need for further fine-tuning.

6.3.3 Discussion

In recent years, the field of diagnostic decision support and therapeutic management support has witnessed significant advances, particularly with the emergence of patient similarity network paradigms (Pai *et al.*, 2019). Researchers have increasingly focused on learning patient profile representations through the development of supervised models and solving constrained optimization problems. Notably, one important distinction from previous studies lies in the utilization of supervised machine learning to determine patient similarity, with only a limited number of articles exploring the use of unsupervised models in research (Brown, 2016; Panahiazar *et al.*, 2015).
While a few studies have attempted to aggregate multiple similarity measures based on various attributes (Pai & Bader, 2018), many of these investigations overlook the importance of specific patient characteristics during the similarity calculation process. Particularly, these methods have not been extensively applied in the context of cancer care, making our research novel in its approach.

In this research, we used two methods to address similarity calculation between two breast cancer patients, the first of which used a unique similarity measure to calculate similarity between patients. The other method involved using various similarity measures for distinct data categories. Finally, we used an expert-made gold standard to find the weights of important variables.

The choice of a structured data model is a paramount step in implementing similarity measures. Having access to a domain ontology, we could have chosen a graph model using the ontology and the entity-attribute-value model to calculate the similarity between patients with graph-based models such as Graph Neural Networks (GNN). Gu *et al.* (2022) has proposed a deep learning framework called "Structure-aware Siamese Graph Neural Networks" (SSGNet) that organizes computerized patient records as graphs and uses GNNs to learn patient similarity. But having access to a limited number of annotated data, we chose to make a simple representation in the form of a table, with for each patient the list of decision variables described before.

The results we obtained make us think that choosing a similarity measure according to the type of variables and having a weighted average of different measures can improve results. Indeed, in Method 1 (use of a single similarity measure), the absence of weighting and the equal importance given to all variables lead to results far from the gold standard. What's more, applying a different measure to each type of variable makes more sense than using the same measure for all types of variables. For example, histological variables, which have a hierarchical tree structure, are

fundamentally different from other types of variables. Using the same method for all variables also makes it difficult to assign weights when considering each observation as a vector with all attribute values.
In addition to improving results, this approach also helps to explain them. Indeed, the weights chosen reflect the importance accorded to the variables. Using this reasoning, we can include expert knowledge when assigning weights (for example, in our case we gave significant weight to the patient's vulnerability variable because experts gave us this information).

Despite our efforts to optimize the single similarity method using deep metric learning, the results were not as promising as initially anticipated. While the Siamese network exhibited improved scores for some patients, reaching up to 3/5 similarity, it encountered significant challenges in accurately identifying similarities for most cases. For instance, patient Pat 41 received a score of 0/5, despite initially achieving a score of 1/5 in the Euclidean similarity calculation. This discrepancy highlights a limitation in the network's ability to generalize and capture the intricated patterns that define patient similarities, moreover, learning new embeddings makes us lose the clinical reality of the data that was processed from BCPSs, which makes it impossible to explain the results obtained. The architecture and training process of the Siamese network may have contributed to these suboptimal outcomes, underscoring the need for further refinement and fine-tuning, or the direct use of the textual BCPSs instead of a structured format extracted from the text. Despite this setback, the deep metric learning approach provides valuable insights (Gu *et al.*, 2022) and may still have the potential for improved performance with additional adjustments and optimizations. Future investigations could explore alternative network architectures, data augmentation techniques, or fine-tuning hyperparameters to enhance the deep learning efficacy in patient similarity determination.

Finally, even if we have obtained numbers for precision and recall, it's crucial to confirm if the method actually works. To do this, we need to show similar patient cases obtained by the algorithms to experts and ask them if the algorithm's suggestions are accurate.

Moreover, it's essential to increase the number of BCPSs used for validation. In our study, the annotation process involved selecting the top 5 most similar patients, which proved to be time-consuming. This required an expert to compare each BCPS from the *similarity evaluation dataset* with the 100 BCPSs in the *similarity learning dataset*. Consequently, we were limited in our ability to annotate more BCPSs for validation. Additionally, involving multiple experts in the annotation process could potentially enhance our results.

6.4 Conclusion

In conclusion, the research work conducted demonstrates the significance of selecting appropriate similarity measures tailored to the type of variables used and effectively weighting these measures. This approach yields favorable results for the accurate detection of similar patients. By incorporating the importance (weight) attached to each variable, we can comprehensively explain the obtained results and integrate expert knowledge in the weighting process. The combination of these strategies not only improves the precision and recall scores of patient similarity methods but also enhances our ability to interpret and validate the outcomes in the context of clinical relevance.

CHAPTER 7

Conclusion

7.1 Summary

This research journey started with the goal of developing a multifaceted computerized decision support tool to assist clinicians in the management of breast cancer patients during multidisciplinary tumor board meetings. With the initial hypothesis that one single system can't fit all patient situations, we wanted to create a CDSS that uses the complexity of a clinical case to provide decision support (guideline-based for non-complex cases and case-based for complex cases).

Throughout the thesis, we addressed various aspects of clinical decision support, including data extraction from unstructured clinical notes, case complexity classification, case-based decision support, and the update of guideline-based decision support.

First of all, despite the emergence of new AI tools to deal with textual input, there are still many tasks that can be automatically performed with structured data. Since patient cases discussed within MTBs are presented as textual breast cancer patient summaries, we developed a rule-based method for the efficient extraction of structured data from BCPSs (chapter 3). The developed pipeline demonstrated strong performance, it was evaluated on 30 textual BCPSs, showing an average F1-score of 0.93 for tumor attributes, 0.8 for side attributes, and 0.85 for patient attributes which show really good performances compared to similar works. These first results emphasized the significance of customization and adaptation to the specificities of BCPSs to achieve higher scores.

Following our hypotheses, we elaborated the decision support modules according to the complexity of clinical cases. We were facing a research problem to choose the best method for BCPS classification with respect to complexity. Even if this is a complex task as there is no definition of breast cancer complexity, our findings revealed that feature extraction using semantic annotators achieved higher results compared to using pre-trained language models. We obtained (89% accuracy when using MLP with semantic annotators compared to 72% accuracy using BERT models and 77% for Word2vec to do the feature extraction on a corpus of 1042 BCPSs annotated as complex and non-complex (80% for training and 20% for test) Classic machine learning methods outperformed transformer models such as BERT in identifying complex cases. This highlights the relevance of considering traditional ML models for this specific task (Chapter 4).

Building upon the understanding of case complexity, we utilized the guideline-based decision support systems of the DESIREE project for non-complex cases. When using the system we faced the problem of updating the system's knowledge base as it was implemented on guidelines of 2016. We implemented a semi-automated method based on real breast cancer cases for identifying practice evolution according to the latest evidence, we used 160 real annotated BCPS that corresponded to non-complex cases. Among them, there were 38 (23%) patient profiles where the recommendation of the system did not comply with the MTB decision and for which we noticed an evolution in practice. This method allowed us to identify update needs in the GL-DSS's knowledge base and add new rules and concepts. Furthermore, the evaluation process supported the fact that guideline-based reasoning is not suitable for complex cases, as the system was non-compliant in 23 out of 34 complex cases (68%), aligning with our initial hypothesis that guideline-based systems are not relevant for managing complex cases (Chapter 5).

Ultimately, regarding decision support for complex cases, we conducted a study to establish a case-based reasoning system for breast cancer patients. Two different approaches were examined for measuring the similarity between patients. The first approach utilized a unique similarity measure that was optimized using deep metric learning, while the second approach was a hybrid method that combined multiple measures and incorporated expert knowledge using weighted averages. Both methods were trained on a corpus of 100 randomly selected BCPSs for which experts manually clustered similar cases. The evaluation was done using the comparison of the top 5 most similar patients for 10 complex BCPSs. The hybrid method yielded encouraging outcomes in determining patient similarity, substantially enhancing the accuracy of assessing similar patients and improving interpretability, achieving a precision rate of 64%. However, deep metric learning methods performed poorly with a precision rate of only 21%, suggesting the need for further improvements to attain the desired results (Chapter 6).

In conclusion, this thesis has provided valuable insights into the development of a clinical decision support system tailored for multidisciplinary tumor board clinicians. By implementing an NLP pipeline that efficiently extracts data from BCPSs and integrates it into a guidelines-based DSS, we have answered our research question on creating an effective system that utilizes clinical notes to provide personalized treatment recommendations based on guidelines.

The complexity prediction algorithm, combined with the results from the guideline-based DSS, addresses the research question on clinician acceptance of guideline-based DSSs in clinical practice. By filtering out non-complex cases and focusing on profiles covered by clinical guidelines, our system offers evidence-based and accurate recommendations, promoting its acceptance among clinicians.

Moreover, the case-based decision support systems, although yet to be formally evaluated, present a promising answer to the question of providing decision support for complex cases. By offering personalized recommendations according to the patient's profile, similarity-based reasoning holds great potential in aiding clinicians faced with intricate clinical scenarios.

Throughout our work, we have adhered to the principle that "one size does not fit all," aligning with our initial objective. By tailoring decision support based on case complexity, we have effectively addressed the main research question of developing a CDSS that assists MTB clinicians in their decision-making process. This approach has demonstrated promising results to enhance collaborative decision-making and ultimately improve patient care. As we continue to refine and evaluate the system, we aspire to contribute significantly to the advancement of clinical decision support in oncology, paving the way for more informed and personalized treatment strategies in the future.

7.2 Limitations

As with any scientific research, this study is not without limitations, which are essential to acknowledge to provide a balanced perspective on the findings and implications. These limitations encompass various aspects of the research, including design, technical, and practical considerations.

7.2.0.1 Design limitations

The research design played a pivotal role in shaping the outcomes of this study. One of the primary design limitations is the sample size and generalizability of the results. The decision support modules' evaluation relied on a limited dataset of breast cancer patient summaries available within the EDS data warehouse. Although the dataset was meticulously annotated, a larger and more diverse dataset from multiple institutions would enhance the generalizability of the developed decision support systems to a broader patient population.

Another design limitation pertains to the focus on breast cancer management. While this focus aligns with the research's specific objectives, it restricts the direct applicability of the developed decision support systems to other cancers, or more generally to other pathologies. Expanding the scope of the study to encompass other cancers would provide a more comprehensive evaluation of the decision support systems' versatility and potential.

7.2.0.2 Technical limitations

Several technical limitations were encountered during the course of this research. One of the main technical challenges was the limited availability of annotated data. Although the datasets were annotated by experts, variations in complexity assessments between different annotators emerged due to the lack of a standardized definition of complexity. Such discrepancies influenced the performance of the complexity classification algorithm and may call into question its reliability.

Additionally, the decision to utilize data from the EDS data warehouse introduced practical constraints and technical challenges. The structuring of breast cancer patient summaries within the data warehouse, often characterized by copy-paste text, required rigorous data cleaning and processing. While efforts were made to address these issues, it remains possible that some noise and inconsistencies persisted in the data, impacting the performance of the decision support modules.

7.2.0.3 Practical limitations

Practical limitations also influenced the research outcomes. One significant practical challenge was the limited time available to work with domain experts including a short number of annotations. The collaboration with clinicians and experts was invaluable for refining the decision support modules and ensuring clinical relevance. However, due to time constraints, the level of expert involvement was constrained, potentially limiting the depth of insights and clinical validation of the developed systems. For instance, we started working on using machine learning methods within the NLP pipeline implemented in chapter 3, but we did not have annotated data to evaluate the method (the machine learning method for NLP we implemented is explained in appendix A).

Moreover, the implementation of guideline-based decision support modules was limited by the environment. The integration of the GL-DSS system into the EDS data warehouse environment, along with the installation of NLP tools, proved to be technically challenging. As a result,

evaluating certain components on a larger scale or real-world data was not feasible within the timeframe of this study.

7.3 Future perspectives

As we conclude this research journey, several promising future research perspectives emerge, offering opportunities to further enhance clinical decision support systems and contribute to improved patient care. These perspectives can be grouped into key areas of focus.

7.3.0.1 Data quality and integration

Improving the quality and structure of clinical notes is essential to overcome challenges related to unstructured data. Replacing the copy-paste format in BCPSs with standardized templates would enhance data retrieval and quality. Additionally, exploring advanced methods for structured data extraction, such as hybrid approaches combining rule-based techniques with machine learning, holds promise for more accurate and efficient data processing.

Seamless integration of decision support systems with hospital information systems is crucial for real-world implementation. This integration would enable direct assessment of the systems' impact on multidisciplinary tumor board meeting workflows, optimizing their support during patient management discussions and decision-making.

7.3.0.2 Validation and user-friendly interface

Validating the decision support systems with expert input and feedback is essential to ensure their clinical relevance and accuracy. Collaborating with healthcare institutions and clinicians outside the research setting would provide valuable external validation and foster wider adoption. Moreover, developing an intuitive graphical user interface and visualization tool for case-based reasoning would enhance the usability and acceptance of the systems in multidisciplinary tumor board meetings.

As the decision support systems move closer to real-world implementation, evaluating their impact on clinical workflow and patient outcomes becomes crucial. Conducting rigorous assessments of the systems' effectiveness in streamlining decision-making processes and improving patient care will be vital to demonstrating their value and benefit to clinical practice.

7.3.0.3 Optimizing similarity and large language model-based methods

Patient similarity-based decision support implemented in chapter 6 offers a promising area for future exploration. To optimize patient similarity assessments, further investigation of deep metric learning techniques can lead to significant enhancements. Additionally, exploring more complex patient representations, such as graph-based approaches, holds the potential for refining the assessment of patient similarity. Similarly, this kind of complex representation of the patient could enhance the algorithm for complexity classification implemented in chapter 4.

Moreover, the emerging trend of utilizing large language models opens up new possibilities for decision support. Integrating these models into the decision-making process could for example replace the manual comparison of guidelines (as we did in chapter 5), enabling a more automated and efficient update of guidelines based on the latest evidence and research findings. This

advancement may streamline the decision-support process, facilitating faster and more accurate clinical recommendations for non-complex cases.

7.3.0.4 Expanding the scope

One crucial avenue for future research lies in extending the scope of the decision support systems beyond breast cancer management. While this thesis primarily focused on breast cancer, adapting and evaluating the systems for other cancer types would provide valuable insights into their potential applicability across different MTBs. Exploring the unique challenges and complexities of various cancers and tailoring the decision support modules accordingly will be paramount to their success.

Extended French summary – Résumé étendu en français

1. Introduction

Ayant dépassé le cancer du poumon en tant que cancer le plus couramment diagnostiqué dans le monde, le cancer du sein constitue une préoccupation majeure pour la santé des femmes. Avec environ 2,3 millions de nouveaux cas diagnostiqués dans le monde, il s'agissait de loin du cancer le plus fréquemment diagnostiqué chez les femmes en 2020 (Sung *et al.*, 2021). Cette même année, le cancer du sein a emporté la vie d'environ 685 000 femmes, représentant une proportion significative des décès par cancer, avec 1 femme sur 6 touchée. D'ici 2040, le nombre de nouveaux cas de cancer du sein diagnostiqués devrait augmenter de plus de 40 %, avec environ 3 millions de cas diagnostiqués chaque année (Arnold *et al.*, 2022).

Réunions de concertation pluridisciplinaire

Dans de nombreux pays, les réunions de concertation pluridisciplinaire (RCP) ont été introduites afin de promouvoir un processus de décision collaboratif dans la prise en charge des patients atteints de cancer. Les experts de diverses spécialités se réunissent pour élaborer le meilleur plan thérapeutique possible pour un patient atteint de cancer. Au cours d'une RCP, les cliniciens examinent l'historique médical du patient dans sa globalité, comprennant les résultats d'imageries et d'autres informations pertinentes permettant d'établir un diagnostic précis et de déterminer le traitement le plus adapté. Les plans théraptiques tiennent compte des besoins et des caractéristiques uniques de chaque patient, tout en tenant compte des recommandations des guides de bonnes pratiques cliniques (GBP).

Bien que des études ont montré que les RCP sont efficaces pour améliorer la conformité des décisions thérapeutiques aux recommandations des GBP (Kesson *et al.*, 2012; van Hoeve *et al.*, 2014; Brar *et al.*, 2014), celles-ci sont remises en question. En effet, des équipes peuvent être touchées par des pénuries de personnel, une augmentation de la charge de travail, un nombre croissant de cas à discuter et une diversité disciplinaire (Soukup *et al.*, 2022; Blayney, 2013; El Saghir *et al.*, 2013).

Néanmoins, les RCP se sont généralisée pour la prise en charge moderne des cancers et jouent un rôle essentiel pour optimiser le pronostic des patients. Aussi, l'amélioration de l'organisation des RCP est un impératif pour garantir un égal accès à tous les patients atteints de cancer et créer

le meilleur environnement possible pour les cliniciens au profit d'une prise en charge globale, singulière et éclairée.

Systèmes d'aide à la décision médicale

Les systèmes d'aide à la décision médicale (SADM) sont considérés comme des outils au potentiel important pour les soins de santé modernes. Ils fournissent une assistance informatisée aux cliniciens dans la prise de décision pour leurs patients. Les SADM sont reconnus pour améliorer la prise en charge des patients atteints de cancer en aidant les médecins à prendre des décisions personnalisées. En délivrant des informations fondées sur des preuves, les SADM peuvent aider les cliniciens à identifier les patients à haut risque, à affiner les diagnostics, à recommander des options de traitement appropriées et à surveiller l'évolution du traitement.

Dans le paradigme de la médecine fondée sur des preuves (Evidence-Based Medicine Working Group, 1992), plusieurs SADM ont été développés et évalués pour promouvoir la prise de décisions cliniques fondées sur des preuves en oncologie. De tels systèmes s'appuient souvent sur les connaissances contenues dans les GBP, qui représentent l'état de l'art. Cependant, la mise en œuvre de SADM basés sur des GBP en pratique clinique pose des défis techniques, notamment l'interopérabilité sémantique avec les dossiers patients informatisés, et la nécessité de valider les données. Aussi, la maintenance des SADM basés sur des GBP et la gestion de multiples GBP présentent d'autres difficultés à résoudre. De plus, les SADM peuvent parfois produire des recommendations non-appropriées, dans les scénarios cliniques ayant des preuves scientifiques limitées (Voigt & Trautwein, 2023).

Il y a quelques années, un nouveau paradigme est apparu dans le domaine médical. La médecine de précision, ou "médecine personnalisée", est une approche innovante visant à adapter la prévention et le traitement des maladies en tenant compte des différences génétiques, environnementales et du mode de vie des individus. L'objectif de la médecine de précision est de cibler les bons traitements, aux bons patients, au bon moment (Gameiro *et al.*, 2018). Les développements récents en intelligence artificielle (IA) sont prometteurs pour révolutionner le domaine de l'oncologie clinique en abordant efficacement de nombreux aspects critiques tout au long du parcours de soins des patients.

De nombreux SADM basés sur l'apprentissage automatique et l'apprentissage profond ont été développés. Ces systèmes ne reposent pas sur des connaissances explicites, mais sur la modélisation de régularités découvertes dans les données disponibles. De grandes quantités de données cliniques sont nécessaires pour la construction de ces modèles et leur réutilisation dans de nouvelles situations similaires. Cependant, même si l'IA présente un énorme potentiel en oncologie clinique, il existe des défis clés à relever pour l'intégrer avec succès dans les soins de routine.

Des recherches récentes (Norgeot *et al.*, 2020; Thompson *et al.*, 2018) ont souligné le nombre limité d'essais prospectifs et d'essais cliniques randomisés pour les modèles d'apprentissage profond, indiquant la nécessité de davantage de validation et de preuves. Des défis tels que les limitations de la quantité de données, l'interprétabilité des modèles et la garantie de la validité clinique, de l'utilité et de l'utilisabilité des modèles d'IA doivent également être surmontés.

Dans l'ensemble, malgré leurs avantages potentiels, l'utilisation en routine clinique de SADM, qu'ils soient basés sur des connaissances ou non, reste limitée (Beauchemin *et al.*, 2019). Il subsiste encore de nombreux travaux de recherche à mener dans ce domaine pour améliorer, élargir et évaluer leur mise en œuvre.

Complexité des cas cliniques

DESIREE est un projet européen dont l'objectif était de créer une plateforme web pour l'aide à la décision pour la prise en charge du cancer du sein primaire. L'un de ses modules d'aide à la décision est le SADM basé sur des GBP (GL-DSS) (Bouaud *et al.*, 2020b), présenté en détail dans la section 2.2.4.1. Dans le cadre de DESIREE, nous avons mené une évaluation du GL-DSS, dont la base de connaissances était basée sur les recommandations françaises de l'AP-HP publiées en 2016.

Grâce à cette évaluation, nous avons identifié des cas où le système n'a pas généré de propositions thérapeutiques pour certains patients ou a recommandé des traitements qui n'ont pas été suivis par les cliniciens de la RCP. Nous avons découvert que ces cas n'étaient pas soit couverts par les GBP, soit présentaient des caractéristiques particulières, nécessitant généralement des discussions pluridisciplinaires approfondies lors des RCP. Après avoir échangé avec des oncologues au sujet de ces profils, les experts ont exprimé que ces types de cas cliniques posent des défis pour leur prise en charge et nous avons qualifié ces profils cliniques de "cas complexes" (Redjdal *et al.*, 2021c). Ces experts ont déclaré que pour ces cas complexes, le rappel de cas similaires avec les décisions prises pour eux serait une aide pour déterminer le plan de soins approprié. Par conséquent, la capitalisation et l'exploitation des cas complexes pourrait constituer le fondement d'une forme alternative d'aide à la décision. Étant entendu que ces cas cliniques des patients puissent présenter différents niveaux de complexité, la reconnaissance des cas complexes pourrait permettre d'accorder plus de temps aux RCP pour discuter ces cas.

Documents cliniques

Les documents cliniques hospitaliers, tels que les comptes rendus d'imagerie ou de pathologies, sont une source d'information précieuse. Il a été estimé que 80 % des données hospitalières sont recueillies sous forme de texte (Raghavan *et al.*, 2014). Cependant, le format de texte libre peut limiter l'usage de ces informations pour les soins cliniques, la recherche et d'autres applications. Une façon de relever ce défi est d'utiliser des techniques d'extraction d'informations (EI) pour structurer automatiquement le contenu des documents cliniques. L'EI consiste à identifier et extraire des éléments spécifiques du texte, tels que les données démographiques des patients, les diagnostics ou les procédures. Ces données structurées peuvent ensuite être utilisées à diverses fins, tels que l'aide à la décision clinique, la recherche ou le codage médicale.

En ce qui concerne la prise en charge du cancer, pendant les RCP, les cliniciens se réfèrent à un document, généralement produit par le médecin en charge du patient avant la RCP. Le médecin rassemble toutes les informations nécessaires pour prendre une décision pour son patient, y compris les antécédents cliniques, les résultats radiologiques, les résultats histologiques, la réponse au traitement, etc., et résume toutes ces informations dans un document textuel. Ce document est partagé au sein de la RCP et complété par la décision de la RCP. Il est appelé "fiche RCP" (F-RCP). C'est un document crucial pour la RCP. Cependant, la F-RCP reste rédigée en langage naturel, et contient de nombreuses abréviations et acronymes, ce qui rend l'utilisation du contenu de la fiche loin d'être directement traitable par un SADM.

Questions de recherche

Globalement, les RCP jouent un rôle crucial dans la prise en charge du cancer en facilitant la prise de décision collaborative par les professionnels de santé. Cependant, leur bénéfice est remis en cause, notamment à cause de l'incidence croissant des cas de cancer du sein et du manque de

temps. Par ailleurs, malgré leur potentiel pour améliorer les RCP, les SADM ne sont pas couramment utilisés dans la gestion des patients atteints de cancer. Cela soulève la question de recherche principale : **"Comment développer un SADM pour assister efficacement les cliniciens de la RCP dans leur processus de prise de décision ?"**

Suite à nos recherches et discussions au sein du projet DESIREE, les cliniciens estiment souvent inutile d'appliquer des SADM basés sur des GBP pour gérer les cas complexes, car ces cas ne sont généralement pas couverts par les GBP. Par conséquent, la question de recherche se pose : **"Que peut-on faire pour que les cliniciens acceptent un SADM basé sur des GBP en tenant compte de la complexité des cas cliniques, et comment pouvons-nous mettre à jour efficacement ces SADM ?"** De plus, **"Comment pouvons-nous aider les cliniciens à gérer les cas complexes lors du processus de prise de décision ?"**

Enfin, les comptes rendus cliniques, tels que les F-RCP, contiennent des informations cruciales pour la prise de décision. Cependant, étant rédigés en langage naturel, les F-RCP présentent des défis pour une utilisation efficace dans les SADM. Ainsi, une autre question de recherche émerge : **"Pouvons-nous créer efficacement un système qui prend les F-RCP en tant qu'entrée, fournissant des recommandations de traitement personnalisées pour les patients ?"**.

Objectifs

L'objectif principal de ce travail peut être résumé dans l'expression : "One size does not fit all !". Nous savons d'après les expériences précédentes, que les SADM basé sur les GBP ont le potentiel d'améliorer la conformité des décisions des RCP aux GBP (Seroussi *et al.*, 2012a), mais nous savons aussi que ces systèmes ont des limites en ce qui concerne les cas cliniques complexes (Redjdal *et al.*, 2021b). Par conséquent, l'objectif de cette thèse est de créer un SADM qui soutienne la prise de décision des RCP de cancer du sein. Suivant notre hypothèse, la démarche retenue est de construire un système qui permette dans un premier lieu d'identifier automatiquement les cas complexes, permettant aux cliniciens de mieux organiser la RCP et de se concentrer sur ces cas complexes qui nécessitent des discussions plus approfondies. Ensuite, au-delà de la reconnaissance des cas complexes, le système de classification de complexité servira de système de triage pour fournir une aide à la décision adéquate.

Les objectifs sont les suivants :

- **Aide à la classification de la complexité :** Comme il n'existe pas de définition de la complexité, nous avons proposé d'utiliser le machine learning pour apprendre la complexité. Un algorithme qui classe les cas de cancer comme complexes ou non complexes a été développé par apprentissage automatique en utilisant différentes techniques d'extraction de caractéristiques à partir des F-RCP (voir le chapitre 4).

- **Aide à la décision pour les cas non complexes :** Comme ces cas sont "simples", nous supposons qu'ils sont correctement couverts par les GBP. Par conséquent, pour ces cas, nous réutilisons le système GL-DSS du projet DESIREE, tel que documenté dans (Bouaud *et al.*, 2020a), mais nous devons l'adapter aux GBP plus récents sur la prise en charge du cancer du sein (voir le chapitre 5).

- **Aide à la décision pour les cas complexes :** Comme ces cas ne sont pas adéquatement couverts par les GBP, une approche d'aide à la décision basée sur des cas similaires est proposée. Le chapitre 6 présente une méthodologie pour détecter les patientes les plus similaires à une patiente donnée. Ce système recommandera des options de traitement aux cliniciens en se

basant sur les décisions prises pour des patientes similaires, reproduisant ainsi le raisonnement médical précédemment utilisé.

Une partie importante de ce travail concerne le traitement des F-RCP en utilisant des techniques de traitement automatique du langage naturel (TAL). En effet, une tâche préliminaire obligatoire consiste à transformer le contenu des F-RCP en un format de données structurées formelles qui permet l'utilisation des SADM.

En réalisant ces objectifs, nous ambitionnons de contribuer au champ de l'aide à la décision en développant un SADM qui optimise le processus de prise de décision au sein des RCP. Par l'intégration de l'aide à la décision basée sur les GBP pour les cas non complexes, et de l'aide à la décision basée sur un raisonnement à partir de cas pour les cas complexes, ainsi que l'utilisation de techniques de TAL pour structurer les données et des techniques d'apprentissage automatique pour la détection de la complexité, ce projet s'efforce de fournir un SADM robuste et efficace, répondant aux besoins des cliniciens participants aux RCP.

Exploitation des données

Dans ce travail, l'objectif était d'utiliser des technologies avancées telles que l'apprentissage automatique et l'apprentissage profond associés à des méthodes d'IA symbolique pour développer un SADM pour les cliniciens de la RCP de sénologie de l'hopital Tenon (Paris, France). Pour garantir la validité et l'approbation éthique du projet, il a été examiné et validé par le comité scientifique et éthique de l'AP-HP (projet CSE200094). Pour nos recherches, nous avons obtenu l'accès aux F-RCP de patients atteints de cancer du sein diagnostiqués entre 2018 et 2022 et traités à l'hôpital Tenon. Ces F-RCP sont accessibles dans l'Entrepôt de Données de Santé (EDS) de l'AP-HP. Pour ce travail, nous avons eu accès à une base de données composée de 11 205 F-RCP, associées à 3 500 patients.

Les jeux de données extraits pour le projet CSE200094 ont été utilisés pour extraire des ensembles de données qui ont servi à l'entraînements et à l'évaluation de chacune des tâches différentes du projet, et peuvent être résumées comme suit :

- Pour la tâche d'extraction de données structurées (SDE) (chapitre 3), nous avons effectué une sélection aléatoire de F-RCP sans duplication, ce qui a donné un ensemble de données de **80 F- RCP**. Parmi elles, nous avons utilisé 50 F-RCP pour développer des règles pour la structuration. Ensuite, nous avons utilisé le reste des F-RCP (30) pour évaluer l'algorithme.

- Parallèlement, nous avons sélectionné **1 048 F-RCP** représentant des patientes discutées entre novembre 2020 et février 2022 pour la tâche d'apprentissage de la complexité (chapitre 4). Un panel d'experts, composé d'un sénior, de deux experts avancés et de trois juniors, a annoté ces F-RCP comme étant complexes ou non complexes. Ce corpus a servi à l'apprentissage supervisé de l'algorithme de détection de la complexité. De cet ensemble de données, 80 % ont été alloués à l'entraînement, et 20 % pour le test.

- En utilisant les annotations de complexité, nous avons également dérivé 160 F-RCP représentant des cas non complexes à partir de l'ensemble d'apprentissage de la complexité. Ce sous-ensemble de données a été utilisé pour évaluer et mettre à jour le SADM basé sur les GBP du projet DESIREE, comme détaillé dans le chapitre 5.

- Enfin, pour la tâche de calcul de la similarité (chapitre 6), nous avons également sélectionné deux sous-ensembles de données à partir de l'ensemble de données d'apprentissage de la

complexité. Le premier sous-ensemble de données comprenait **100 F-RCP (50 complexes et 50 non complexes)** représentant des patientes dans la même situation clinique (patientes ayant subi une chirurgie sans traitement néoadjuvant, désignées ci-après comme patientes dans le scénario D). Un expert avancé a regroupé cet ensemble de données en clusters de patientes similaires. Puis, le deuxième sous-ensemble de données composé de 10 cas complexes dans le scénario D a été utilisé pour évaluer l'algorithme de calcul de la similarité. Un expert avancé a calculé, pour chacune des 10 F-RCP, les 5 F-RCP les plus similaires à partir de l'ensemble des 100 F-RCP groupées en clusters.

Pour une représentation visuelle de l'organisation des ensembles de données tout au long de la thèse, veuillez-vous référer à la Figure 7.1.

2. Extraction de données structurées

Dans le domaine de la santé et de la recherche médicale, les dossiers patients informatisés (DPI) jouent un rôle crucial en tant que sources d'informations médicales. Les DPI contiennent des données essentielles pour la prise en charge des patients. Cependant, le manque de normalisation des DPI pose des défis en matière de réutilisation efficace de leur contenu pour la recherche. La plupart des informations contenues dans un DPI sont fournies sous forme de texte (80 %) (Raghavan *et al.*, 2014). Dans cette partie de la thèse, nous nous sommes concentrés sur l'extraction de données structurées à partir des F-RCP issues de l'EDS de l'AP-HP.

En s'inspirant du modèle d'information du projet DESIREE, basé sur le modèle entité-attribut-valeur (EAV) et représenté dans une ontologie nommée BCKM (Breast Cancer Knowledge Model), nous avons créé une pipeline intégrant des techniques basées sur des règles ou patrons syntaxiques pour annoter les expressions faisant référence à des éléments de données structurées.

Fiches RCP

Comme indiqué dans l'introduction, ce projet a été approuvé par le comité d'éthique institutionnel de l'AP-HP (CSE 200094). Nous avons eu accès à un échantillon de plus de 11 000 F-RCP. Une F-RCP typique fournit un portrait de la patiente fournissant toutes les informations pertinentes dont les cliniciens ont besoin pour prendre une décision. Les informations sont le plus souvent organisées selon l'ordre suivant :

- Informations personnelles (disponibles de manière anonyme dans l'entrepôt de données)
- Données biométriques
- Raison de la présentation
- Antécédents personnels (antécédents médicaux et chirurgicaux, traitements suivis et allergies.)
- Antécédents familiaux
- Histoire de la maladie
- Examen clinique
- Résultats de la radiologie (Mammographie, Échographie, IRM, etc.)

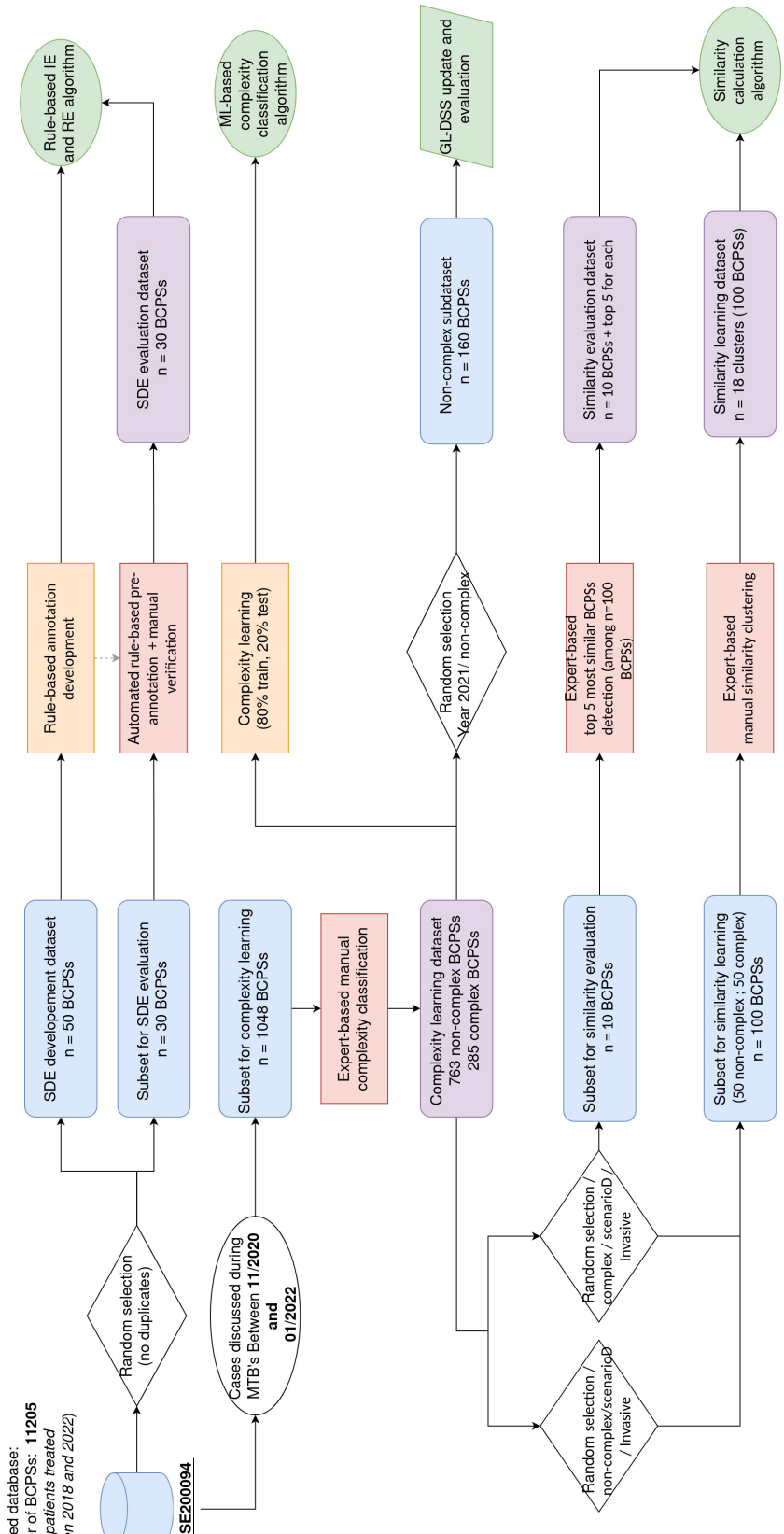


Figure 7.1: Diagramme de type consort illustrant l'utilisation des ensembles de données. *BCPS : Breast Cancer Patient Summary (F-RCP); IE : Information Extraction ; RE : Relation Extraction ; SDE : Structured Data Extraction ; ML : Machine Learning ; GL-DSS : système d'aide à la décision basé sur les GBP du projet DESIREE ; Scénario D : Patients ayant subi une intervention chirurgicale sans traitement néoadjuvant.*

- Résultats de la biopsie
- Classification TNM
- Réponse au traitement néoadjuvant (le cas échéant)
- Résultats de l'anatomopathologie (si une chirurgie préalable a été réalisée)
- Proposition de traitement

Schéma d'annotation

Nous avons organisé la base de données cible en fonction des composants du modèle BCKM (modèle EAV) de DESIREE Bouaud *et al.* (2020b). Les caractéristiques pertinentes à extraire ont été le résultat de discussions avec les médecins de la RCP. Le schéma d'annotation lui-même a été le résultat de plusieurs itérations avec les experts (6 séances d'annotation). Nous conservons les mêmes entités principales que dans le BCKM : le patient, le coté mammaire et la tumeur. Les attributs pour chaque entité ainsi que leurs valeurs potentielles sont décrits dans le tableau 3.2. En plus de ces 3 entités principales, nous avons également extrait les concepts liés aux procédures diagnostiques (IRM, échographie, etc.) et aux procédures thérapeutiques (chirurgie, chimiothérapie, etc.) comme présenté dans le tableau 3.1.

Pour l'annotation, nous avons utilisé l'outil d'annotation BRAT (Stenetorp *et al.*, 2012b). La figure 7.2 montre la version annotée par les experts de la figure 3.1.

Dans l'outil BRAT, seuls 4 types de mentions peuvent être utilisées, à savoir : entité, attribut, événement et relation. Compte tenu de ce modèle, nous avons dû adapter le schéma d'annotation pour qu'il s'y intègre. Par conséquent, les caractéristiques des patientes sont annotées en tant qu'entités, et leurs valeurs sont exprimées sous forme d'attributs.

De plus, BRAT ne permet pas à l'utilisateur d'ajouter une valeur textuelle ou un entier pour un attribut. Toutes les valeurs possibles d'un attribut doivent être placées dans le fichier de configuration de l'annotation. Par conséquent, dans le schéma d'annotation, les attributs ayant des valeurs numériques (ex : la taille de la tumeur) ou des valeurs textuelles (ex : les comorbidités), sont exprimés sous forme de texte annoté. Pour les attributs ayant des valeurs hiérarchiques et booléennes, la valeur de chaque attribut est exprimée à l'aide d'une mention d'attribut.

En plus des entités déjà citées, nous avons également développé des algorithmes basés sur les expressions régulières pour la détection des informations contextuelles dans le texte. Dans BRAT, nous exprimons ces informations contextuelles sous forme de mentions d'attributs, qui peuvent être appliquées à diverses entités pour détecter les entités négatives, hypothétiques, liées à la famille, les préférences du patient et les antécédents.

Enfin, nous avons développé des algorithmes basés sur des règles pour la détection des relations entre les entités. Étant donné que BRAT n'a pas été initialement conçu pour annoter de longues relations multilignes, nous avons essayé d'utiliser le moins de mentions de relation possibles :

- **Has_side :** Cette relation exprime la relation entre une entité de lésion et une entité de côté. Pour éviter des relations multilignes, nous exprimons cette relation en utilisant la mention d'attribut (gauche, droite ou bilatérale) qui peut être appliquée à n'importe quelle entité de lésion ou de type histologique.

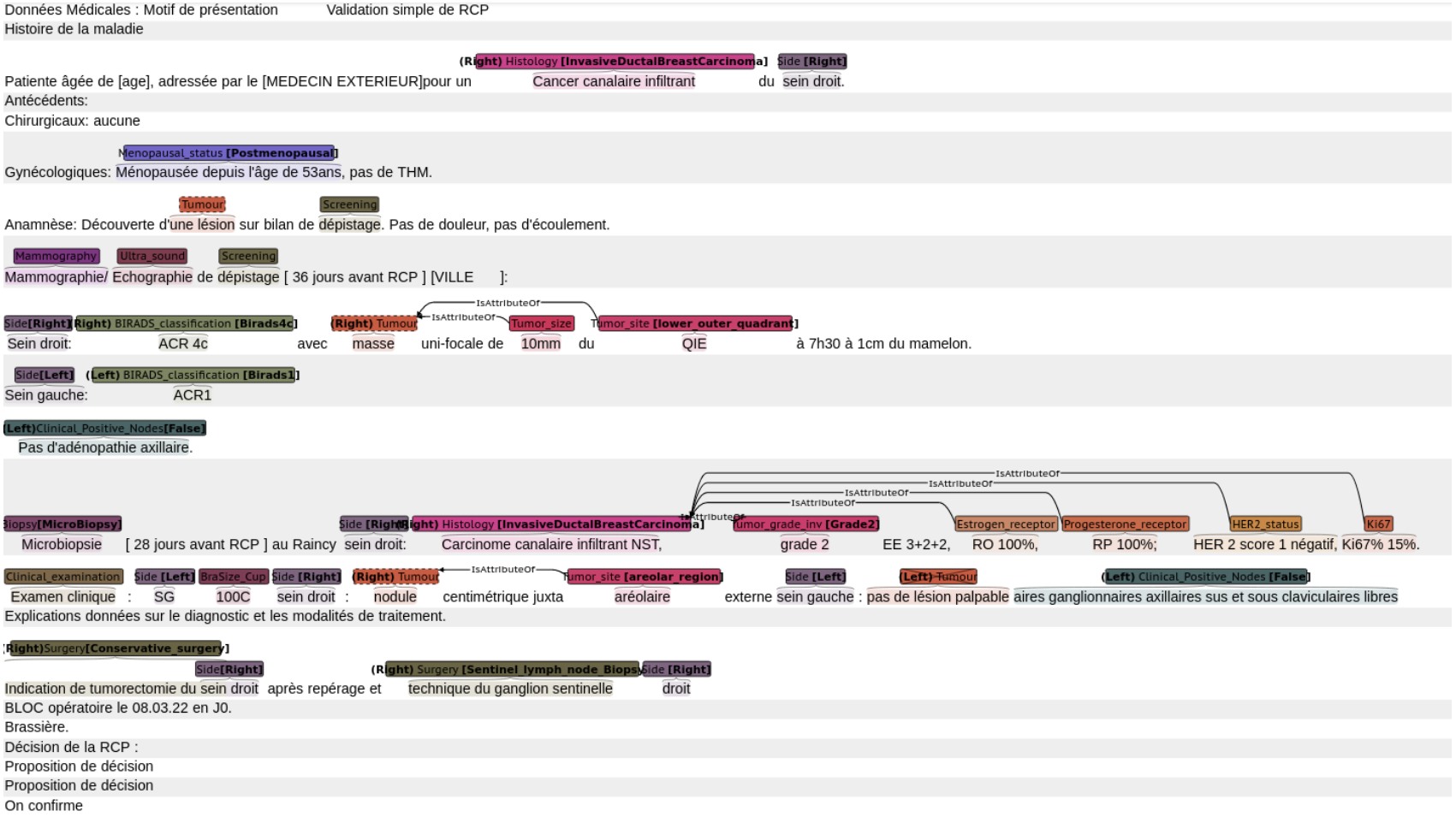


Figure 7.2: Version annotée de la Figure 3.1.

- **Is_attribute_of :** Cette relation est utilisée pour relier les attributs d'une entité à l'entité elle-même. Dans ce travail, nous devons exprimer la relation entre les attributs de côté mammaire et leur entité (quel côté mammaire) ainsi que la relation entre les attributs de lésion et l'entité lésion (quelle lésion). Les attributs de côté pouvant être liés au côté gauche et/ou droit, nous utilisons la même méthodologie que celle utilisée pour exprimer la relation *Has_side*. Pour les attributs de tumeur, nous utilisons un nom de relation "Is_Attribute_of" pour exprimer le lien entre un attribut de tumeur et sa tumeur. Même si nous rencontrons encore des problèmes avec les relations multilignes, nous remarquons que les attributs de tumeur sont généralement exprimés juste après la mention de l'entité de tumeur.

Pipeline d'extraction de données structurées

L'approche proposée a été d'utiliser des règles spécifiques au domaine et des expressions régulières pour identifier et baliser les entités spécifiques mentionnées dans le schéma d'annotation dans le texte. La figure 7.3 illustre tout le processus.

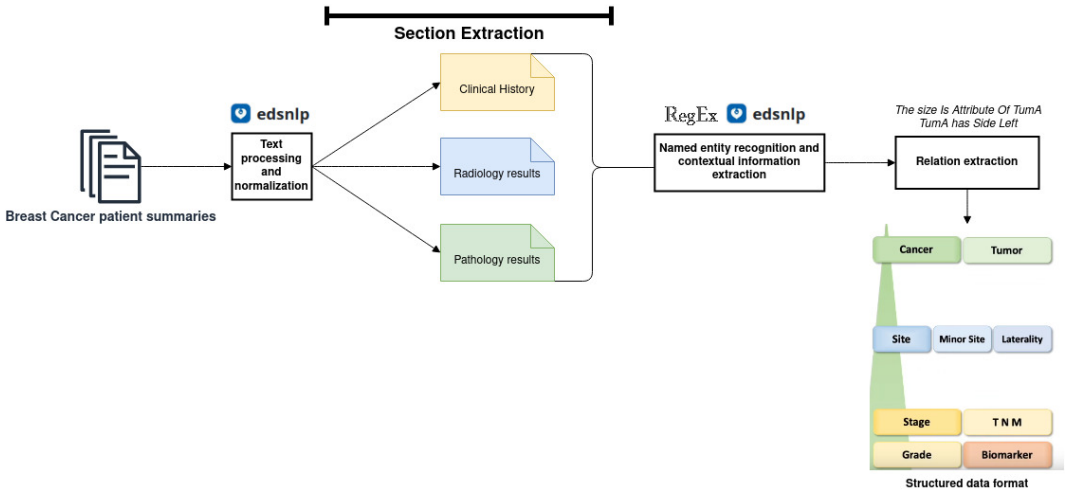


Figure 7.3: Pipeline d'extraction de données strcuturées

La chaîne de traitement commence par normaliser le texte en utilisant les composants de l'outil **EDS-NLP** (Dura *et al.*, 2022b), puis un module détectant les différentes sections est utilisé afin de diviser le texte en sections qui seront traitées individuellement (ex : la section IRM, ou la section histoire de la maladie, etc.). Ensuite, pour chaque section, un algorithme d'extraction d'entités nommées (NER) basé sur des règles est utilisé. Dans cet algorithme, nous avons utilisé les composantes d'EDS-NLP pour extraire les dates, les médicaments et les maladies. Nous avons également utilisé **Clarity NLP** (Georgia_Research_Institute, 2018) pour extraire les entités de taille, de distance et la classification TNM. En ce qui concerne le reste des attributs, nous avons développé des règles et des expressions régulières personnalisées.

Une fois l'algorithme NER exécuté, les informations contextuelles telles que les entités négatives, hypothétiques, liées à la famille, les préférences du patient et les entités antécédents sont identifiées en utilisant des expressions régulières, en plus des composants de négation d'EDS-NLP. Enfin, une fois que l'extraction d'entités nommées et de contexte est faite, nous utilisons un algorithme à base de règles pour extraire les relations.

Le processus d'extraction des relations dans les F-RCP comprend les étapes suivantes :

- Extraction de la relation "Has_side" : Des règles spécifiques ont été utilisées pour établir cette relation. L'idée est de séparer chaque section en phrases et que chaque entité tumeur soit liée au côté mammaire mentionné dans la même phrase (par exemple, tumeur du QSE du sein gauche. La tumeur ici a pour côté mammaire le côté gauche).

- Relation "Is_attribute_of" : De même, ici on considère qu'un attribut de côté est relié au côté si les deux entités sont dans la même phrase ou s'il n'est mentionné qu'un seul côté mammaire dans une section. Concernant les attributs de tumeur, tous les attributs qui suivent une mention de tumeur sont reliés à la tumeur, jusqu'à l'annotation d'une nouvelle entité tumeur ou d'une entité côté mammaire.

Pour évaluer le système, nous avons utilisé des F-RCP de patientes atteintes de cancer du sein annotées manuellement par un expert avec l'outil BRAT. L'évaluation a porté sur trois aspects : les mentions d'entités, les mentions d'attributs et les mentions de relations. Toutes les mentions ont été évaluées en comparant les mentions extraites par le système aux mentions annotées manuellement, en utilisant les mesures de précision, de rappel et de score F1.

Résultats

L'algorithme d'annotation automatique dans sa globalité a été développé en utilisant de petits échantillons de F-RCP, qui ont été discutés de manière itérative avec des experts du domaine pour affiner le schéma d'annotation et ajouter de nouvelles règles pour capturer de nouvelles entités. Après six sessions d'annotation manuelle avec 2 experts avancés (50 F-RCP), nous avons identifié tous les attributs et entités pertinents à extraire. L'algorithme a été évalué sur un corpus annoté manuellement de 30 résumés de patientes atteintes de cancer du sein.

Les résultats de l'évaluation présentés dans le tableau 3.4, montrent l'efficacité de l'algorithme dans l'extraction des entités et de leurs attributs. L'évaluation a été réalisée sur l'extraction d'attributs et de valeurs pour différentes entités, notamment "Patient", "Côté", "Lésion", "Procédures de diagnostic" et "Procédures de traitement".

- Pour l'entité "Patient", les résultats indiquent des performances élevées, notamment pour les attributs "Mutation génétique" et "Statut ménopausique" (score F1 de 0,94). Certaines entités ont des performances légèrement inférieures, comme "BraSize Cup" et "BreastCancer- Relapse" (scores F1 de 0,78 et 0,76). Nous avons constaté un faible rappel pour "comorbidité" (0,64) en raison de l'annotation par l'expert de comorbidités non prises en compte lors de la construction des règles.

- Concernant l'entité "Côté mammaire", nous avons remarqué une excellente reconnaissance pour les attributs "N Status" et "Classification BIRADS" (scores F1 supérieurs à 0,97). En revanche, l'attribut "Confirmed Positive Nodes," faisant référence au statut des ganglions après une chirurgie, a obtenu une performance très faible (score F1 de 0,29). Cela est dû au fait que cet attribut est souvent exprimé de manière hétérogène.

- Enfin, l'entité "Tumeur" est l'entité ayant obtenu les meilleurs résultats, avec un score F1 supérieur à 0,9 pour tous les attributs de cette entité.

En ce qui concerne l'extraction d'informations contextuelles, l'analyse de la performance du processus NLP est décrite dans le tableau 3.5. Nous pouvons constater des performances variables

pour les attributs contextuels. "Hypothétique" a affiché la meilleure performance, avec un score F1 élevé de 0,88. La "Négation" a également montré des performances notables, avec un score F1 de 0,74. Cependant, les attributs "Famille" et "Antécédent" ont montré des résultats relativement modestes, avec des scores F1 de 0,63 et 0,51, ce qui indique des marges d'amélioration dans leur reconnaissance. De plus, l'attribut "Préférences du patient" a été mentionné 6 fois dans les annotations des experts sans être repéré par l'algorithme.

Par ailleurs, le tableau 3.6 présente la performance de l'algorithme d'extraction de relations. L'analyse révèle que le type de relation "IsAttributeOf" a démontré d'excellentes performances, avec un score F1 à 0,90. Pour les relations "Côté Droit" et "Côté Gauche", le modèle a montré des performances respectables (0,79 et 0,83 de score F1). Cependant, la relation qui exprime une tumeur "Bilatéral", a obtenu des résultats modestes avec un score F1 de 0,25 (cas assez rare avec seulement 23 mentions).

Discussion et conclusion

Dans cette étude, l'objectif était d'évaluer l'efficacité d'une méthode basée sur des règles pour extraire des données structurées à partir de documents cliniques. Les performances de la méthode se sont avérées satisfaisantes, avec une précision et un rappel moyen d'environ 0,81 et 0,84 respectivement. L'attention a été portée sur les attributs pour lesquels des règles précises ont été élaborées, obtenant ainsi un score F1 moyen d'environ 0,93 pour l'entité "Lésion". Cela souligne l'importance de developper des règles spécifiques pour certains attributs afin d'améliorer les performances globales du processus.

L'évaluation a permis la découverte d'attributs jusque-là non reconnus, soulignant la nature dynamique des données cliniques et la nécessité d'accepter de nouveaux attributs. Certaines performances suboptimales ont été observées pour des attributs complexes comme le statut des ganglions, suggérant la possibilité d'utiliser des approches plus flexibles, comme l'apprentissage profond. L'efficacité de la méthode basée sur des règles a également été influencée par la qualité des données. Les F-RCP présentaient une grande variation de style et de contenu selon les rédacteurs, ce qui a posé des défis pour l'algorithme basé sur des règles. De plus, la présence d'abréviations a ajouté des ambiguïtés.

Comparativement à d'autres travaux sur des documents en français (Schiappa et al., 2022), cette méthode a démontré des performances compétitives, voire supérieures dans certains cas, concernant les attributs extraits. L'avenir pourrait consister en une approche hybride combinant des règles avec des techniques d'apprentissage profond pour capturer les subtilités contextuelles présentes dans les notes cliniques.

En conclusion, cette étude a montré que l'extraction de données à partir de documents cliniques à l'aide de règles est efficace, mais elle soulève des défis liés à la qualité des données et à la complexité des attributs. Une approche hybride, incluant de l'apprentissage statistique, pourrait offrir de meilleures performances à l'avenir.

3. Apprentissage de la complexité

Comme mentionné dans l'introduction, les cas cliniques des patients peuvent varier en complexité, et il n'existe pas de définition a priori de la complexité du cancer du sein (Soukup *et al.*, 2019). Afin de comprendre et de prédire la complexité des cas cliniques des patients, nous avons exploré différentes approches d'apprentissage automatique ("machine learning").

Classification des textes dans le domaine de la santé.

La classification de documents médicaux est un sous-domaine spécifique de la classification de textes en général. Les algorithmes d'apprentissage automatique supervisés en TAL ont été utilisés avec succès, par exemple les SVM (Support Vector Machine ou Machine à vecteurs de support) et l'analyse discriminante linéaire (LDA) ont obtenu des résultats satisfaisants pour des tâches telles que la classification des documents textuels. Cependant, ces méthodes requièrent une sélection manuelle des caractéristiques (« features »), ce qui peut être difficile et chronophage, conduisant à une représentation limitée des textes.

Les avancées récentes en TAL, telles que Word2Vec, FastText et BERT, ont réduit la nécessité de l'ingénierie manuelle des caractéristiques. Les modèles de langage pré-entraînés tels que BERT et RoBERTa ont été affinés (« fine tunning ») pour des tâches de santé, fournissant de bons résultats (Li *et al.*, 2022). Ces modèles peuvent s'adapter à différents types de données textuelles, les rendant utiles en santé, où le langage peut varier considérablement.

Dans cette partie, nous avons utilisé un corpus de F-RCP annotées par leur complexité pour les classer en fonction de celle-ci, en comparant deux méthodes d'extraction de caractéristiques à partir des textes : (1) en utilisant des annotateurs sémantiques et (2) en utilisant des modèles de langage pré-entraînés.

Annotation des données par des experts

Entre novembre 2020 et janvier 2022, des experts de la RCP de sénologique de l'Hôpital Tenon ont régulièrement annoté les cas des patients comme étant complexes ou non complexes. Ce processus d'annotation a recueilli les raisons de la complexité (quand la F-RCP était annotée « complexe ») pour mettre en œuvre une méthode basée sur des règles pour prédire la complexité et établir une définition formelle locale de la complexité des cas.

Apprentissage de la complexité à l'aide d'annotateurs sémantiques automatiques

Nous avons utilisé des annotateurs sémantiques pour extraire des données structurées à partir de notes cliniques (voir figure 7.4).

ECMT et **MetaMap** ont été employés pour annoter du texte en français (ECMT) et l'anglais (MetaMap). Ces annotateurs ont été utilisés pour extraire des concepts cliniques des textes, y compris les concepts de l'UMLS.

Classification de la complexité basée sur des règles : Pour classer la complexité, nous avons créé un ensemble de règles basées sur les concepts extraits par les annotateurs. Lorsqu'au moins un concept lié à la complexité était présent dans une F-RCP, le cas était considéré comme complexe. La méthode basée sur des règles a montré certaines limites en raison de la variété des facteurs de complexité qui ne sont peut-être pas entièrement couverts par des concepts prédéfinis.

Prédiction de la complexité basée sur l'Apprentissage Automatique : D'abord, nous avons converti les F-RCP en vecteurs de caractéristiques représentant des concepts cliniques extrait par les annotateurs. Ensuite, des modèles d'apprentissage automatique, dont XGBoost et MLP, ont été entraînés sur les données annotées en utilisant une stratégie de validation croisée. Les modèles ont été évalués en utilisant la précision, le rappel, le score F1, la courbe ROC et la courbe PR (voir figure 4.2).

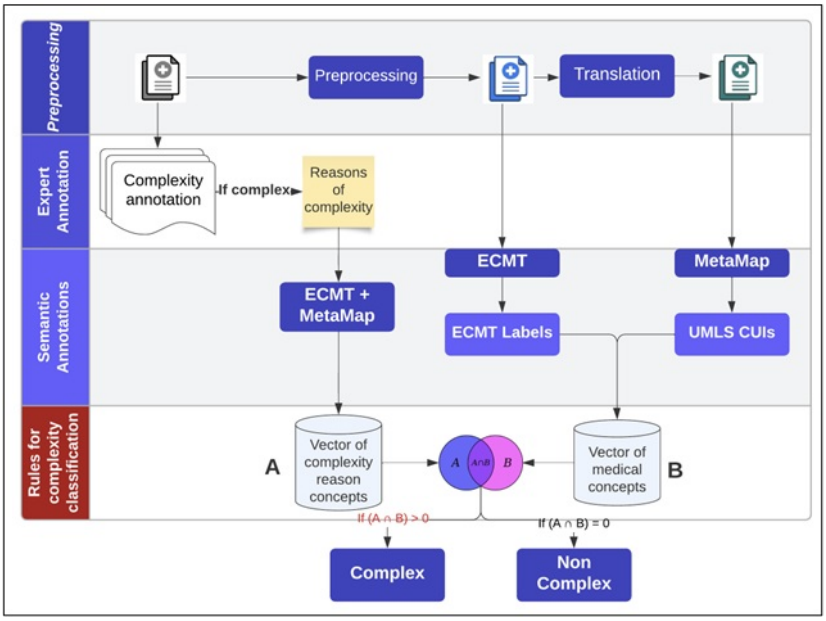


Figure 7.4: Extraction de concepts cliniques et classification basée sur des règles

Apprentissage de la complexité à l'aide de modèles de langage pré-entraînés.

Nous avons exploré l'utilisation de modèles de langage basés sur des architectures de transformateurs (transformers). Cependant, les modèles de transformateur ont posé un défi en raison de leurs limitations en termes de « tokens » acceptés en entrée (512 « tokens »), ce qui a entraîné deux stratégies : la troncature et le fractionnement. Nous avons utilisé un modèle BERT entraîné sur des documents de l'EDS de l'AP-HP pour ces méthodes.

Apprentissage de la complexité à l'aide de modèles statiques

En plus des modèles de transformateurs, nous avons examiné l'efficacité de méthodes antérieures, telles que Word2Vec et GloVe. Ces méthodes ont utilisé des plongements lexicaux de mots pré-entraînées pour capturer les relations sémantiques au sein des textes cliniques.

Résultats et discussion

Nous avons mené l'étude sur un échantillon de 1 048 BCPS, comprenant 763 cas non complexes et 285 cas complexes. Les résultats (tableau 4.1) montrent que la méthode basée sur l'usage des annotateurs sémantiques pour l'extraction de caractéristiques surpasse les méthodes basées sur les modèles de transformateurs et les modèles de langage statique. BERT, a eu du mal à prédire les cas complexes. Word2Vec et GloVe ont fourni une précision compétitive et ont surpassé les méthodes basées sur BERT.

Parmi plusieurs algorithmes d'apprentissage automatique testés sur des F-RCP (Redjdal *et al.*, 2022b), le modèle MLP a obtenu les meilleures performances, suivi de XGBoost. La méthode basée sur des règles, a montré des limites dans la capture de toutes les raisons de la complexité. Des recherches futures pourraient impliquer l'amélioration de la méthode basée sur des règles en utilisant les données structurées issue de la première partie.

Conclusion

En conclusion, sur notre corpus de F-RCP, les modèles d'apprentissage automatique classiques, en particulier le MLP, ont démontré une efficacité supérieure à celle des modèles de transformateurs et aux méthodes basées sur des règles pour la classification de la complexité du cancer du sein. L'utilisation d'annotateurs sémantiques pour extraire les caractéristiques cliniques à partir des F-RCP s'est révélée être une méthode prometteuse. Des perspectives d'amélioration des résultats sont possibles en augmentant le corpus d'entraînement et en utilisant les algorithmes implémentés en première partie.

4. Mise à jour de la base de connaissances du projet DESIREE

Maintenant que nous disposons d'algorithmes pour extraire les données structurées et classifier les patients en complexes et non complexes, nous pouvons passer à l'aide à la décision.

Les SADM jouent un rôle essentiel dans l'amélioration de la qualité des soins de santé en promouvant la conformité des décisions aux guides de bonnes pratiques. Cela réduit les erreurs et assure la conformité aux pratiques recommandées. Le maintien à jour des bases de connaissances est crucial pour maintenir la fiabilité des SADM. Cependant, les mises à jour manuelles sont souvent laborieuses et coûteuses, et les méthodes automatiques de comparaison des textes des GBP ont leurs limites. Dans cette partie, notre objectif est d'identifier l'évolution des connaissances en matière de traitement du cancer du sein de manière semi-automatique. Cela se fait en examinant la corrélation entre les décisions prises par les cliniciens en RC) et les recommandations du SADM pour un profil de patient.

Comme mentionné en introduction, dans cette étude, nous avons utilisé le système GL-DSS, développé pour le projet DESIREE et basé sur les GBP de l'AP-HP de 2016 pour le cancer du sein. Pour mettre à jour la base de connaissances du GL-DSS, nous avons comparé les décisions de la RCP aux recommandations du GL-DSS. Les cas où les décisions de la RCP différaient des recommandations du GL-DSS ont été analysés plus en détail en utilisant les dernières recommandations du SENORIF, publiées en 2021.

Méthode proposée

Comme illustrée dans la Figure 7.5, la méthode proposée comprend quatre étapes. L'hypothèse sous-jacente est que les cas pour lesquels les décisions de la RCP sont conformes aux recommandations de 2016 du GL-DSS, sont des cas pour lesquels il n'y a pas d'évolution des pratiques :

1. **Structuration des données** : Un corpus de comptes rendus de RCP (F-RCP) représentant des cas cliniques non complexes de l'année 2021 est converti en un format structuré en utilisant l'algorithme développé lors de la première partie et est mappé sur l'ontologie BCKM. Cela garantit que les données peuvent être traitées par le GL-DSS.

2. **Comparaison avec les recommandations du GL-DSS** : Les décisions des cliniciens en RCP (prises en 2021) sont comparées aux recommandations générées par le GL-DSS (basé sur les GBP de l'AP-HP de 2016). Les cas où les décisions de la RCP sont incluses dans les recommandations du GL-DSS sont notés comme conformes aux pratiques existantes.

3. **Comparaison avec les recommandations du SENORIF** : Les cas où les décisions de la RCP ne correspondent pas aux recommandations du GL-DSS déclenchent une révision manuelle

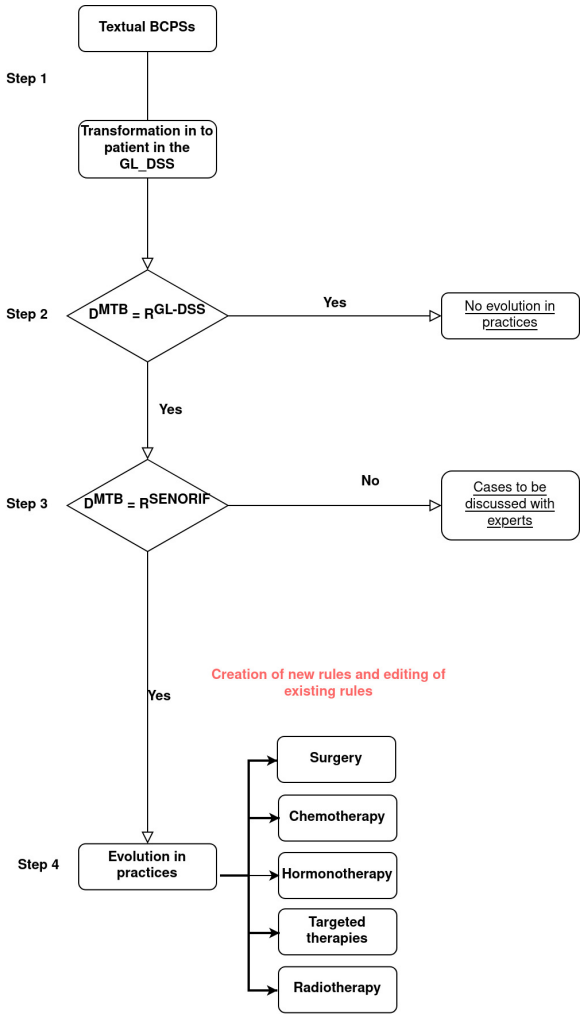


Figure 7.5: Pipeline utilisé pour la mise à jour de la base de connaissances. D^{MTB} = Décision de la RCP ; $R^{GL\text{-}DSS}$ = Recommandations GL-DSS APHP-2016 ; $R^{SENORIF}$ = Recommandations SENORIF-2021.

des recommandations pour le profil du patient dans le SENORIF. Nous catégorisons les cas en deux groupes : ceux où les décisions de la RCP correspondent aux recommandations du SENORIF, indiquant une évolution des pratiques, et ceux qui ne correspondent pas, nécessitant une discussion supplémentaire avec les experts.

4. **Identification des mises à jour de la base de connaissances** : Des mises à jour sont apportées à la base de connaissances du GL-DSS en tenant compte des cas où la décision de la RCP diffère de la recommandation du GL-DSS APHP-2016 et est incluse dans les recommandations du SENORIF 2021. Cela inclut la modification de règles existantes, l'ajout de nouvelles règles, la suppression de règles obsolètes et la correction de problèmes. L'ontologie BCKM est également mise à jour pour inclure de nouveaux concepts.

Résultats et discussion

1. **Structuration des données** : : Étant donné que l'algorithme d'extraction de données structurées a été conçu selon le modèle du BCKM, le mappage a été facile et direct. Une table de mappage entre les variables extraites et les concepts dans l'ontologie est disponible en annexe B.

2. **Comparaison avec les recommandations du GL-DSS** : Sur 160 F-RCP non complexes, les décisions de la RCP en 2021 étaient en accord avec les recommandations du GL-DSS APHP-2016 dans 76 % des cas, suggérant une absence de changement dans les pratiques. Dans les 24 % restants où elles étaient différentes, ces divergences suggèraient une potentielle évolution de pratiques, concernant principalement les interventions chirurgicales (51 %) et la chimiothérapie (41 %).

3. **Comparaison avec les recommandations SENORIF** : Sur les 39 F-RCP (24 %) avec des divergences entre les recommandations du GL-DSS APHP-2016 et les décisions de la RCP, 38 étaient en accord avec les recommandations SENORIF. Cette concordance souligne l'évolution des connaissances, incitant à des mises à jour dans le GL-DSS. Le seul cas non conforme a été revu par des experts et a finalement été reclassé comme cas complexe.

4. **Identification des mises à jour de la base de connaissances** : L'analyse a conduit à 18 modifications de règles, 9 suppressions de règles et l'ajout de 21 nouvelles règles dans la base de connaissances du GL-DSS.

Conclusion

En conclusion, cette étude souligne la nécessité impérative de mises à jour continues de la base de connaissances des SADM. La méthode proposée est nouvelle et vise à utiliser les données de la vie réelle pour détecter les mises à jour d'une base de connaissances d'un SADM. Le GL-DSS a démontré un grand potentiel pour les cas non complexes, mettant en lumière le potentiel des systèmes basés sur la connaissance pour soutenir les professionnels de la santé et améliorer les soins aux patients lorsqu'ils sont maintenus à jour.

5. Aide à la decision basé sur les cas pour les cas complexes (case-based)

Au cours des dernières années, le domaine d'aide à la décision diagnostique et thérapeutique a connu des progrès significatifs, notamment avec l'émergence des paradigmes de réseaux de similarité des patients (Pai *et al.*, 2019). Les chercheurs se sont de plus en plus concentrés sur l'apprentissage des représentations de profils de patients par le développement de modèles supervisés. Notamment, une distinction importante par rapport aux études précédentes réside dans l'utilisation d'apprentissage automatique supervisé pour déterminer la similarité des patients, avec seulement un nombre limité d'articles explorant l'utilisation de modèles non supervisés dans la recherche (Brown, 2016; Panahiazar *et al.*, 2015).

Alors que quelques études ont tenté d'agréger plusieurs mesures de similarité en fonction de diverses attributs(Pai & Bader, 2018), de nombreuses investigations ne tiennent pas compte de l'importance de caractéristiques spécifiques des patients au cours du processus de calcul de similarité. En particulier, ces méthodes n'ont pas été largement appliquées dans le contexte des soins du cancer.

Dans cette partie, notre objectif a été de créer une mesure de similarité, afin de calculer la similarité des patients à un patient atteint de cancer du sein, et de créer un SADM basé sur les cas pour l'aide à la décisions des cas complexes. Nous avons comparé 2 méthodes, l'usage d'une mesure de similarité unique pour tous les attributs et l'usage d'une somme pondérée de plusieurs mesures en fonction du type des attributs.

Matériel et méthodes

Dans cette recherche, nous avons utilisé deux méthodes pour calculer la similarité entre deux patientes atteintes du cancer du sein, dont la première utilisait une mesure de similarité unique (Cosie, Euclidienne ou Jaccard) pour calculer la similarité entre les patientes. L'autre méthode consistait à utiliser diverses mesures de similarité pour des catégories de données distinctes. Enfin, nous avons utilisé un « gold standard » produit par les experts pour apprendre les mesures de similarité et un autre pour évaluer le système.

Le travail a été divisé en plusieurs étapes :

1. **Construction du "gold standard"** : En collaboration avec des experts en oncologie, nous avons créé un jeu de données pour évaluer et entrainer les algorithmes. D'abord, les experts ont analysé 100 F-RCP sélectionnées aléatoirement. Puis, les F-RCP ont été groupées en « clusters » en fonction des caractéristiques cliniques des patientes, formant ainsi un ensemble d'entraînements appelé « Clustering Gold Standard ».

   En outre, un échantillon de 10 F-RCP a été extrait des cas annotés complexes, en excluant les 100 F-RCP précédemment sélectionnées. Les experts ont ensuite choisi, parmi les 100 F-RCP d'entraînements, les 5 F-RCP les plus similaires à chacune des 10 F-RCP, formant ainsi un ensemble d'évaluation baptisé "Top 5 Gold Standard". Ces ensembles de données sont destinés à évaluer les performances des algorithmes.

2. **Construction d'un jeu de données structurés** : En concertation avec les experts, nous avons identifié les attributs les plus pertinents pour chaque entité clé, déterminant également les valeurs potentielles associées à ces attributs. Cette étape a abouti à une liste complète de caractéristiques, à prendre en considération pour la détection de patientes similaires. Pour garantir une comparaison pertinente, nous avons séparé les données en fonction des scénarios de prise en charge des patientes, qui variaient au cours de leur parcours de soins. En outre, nous avons distingué les patientes atteintes d'un cancer du sein invasif de celles atteintes de cancer du sein « in situ », car ces deux catégories présentent des caractéristiques différentes. Finalement, nous avons un jeu de données représentant les 100 patientes du corpus d'entrainement sous la forme d'un tableau représentant des patients en scenario D (prise en charge après chirurgie sans traitement néoadjuvant) et contenant la valeur pour chacun des 27 attributs sélectionnés par les experts pour ce scenario.

3. **Construction d'une mesure générique de similarité** Après une revue de la littérature, nous avons choisi différentes mesures pour évaluer la similarité des cas, que nous avons appliquées de 2 manières :

   - Dans la 1ère méthode, une seule mesure a été utilisée pour tous les attributs, avec normalisation des variables pour assurer une échelle commune.
   - Dans la deuxième méthode, les variables ont été regroupées en fonction de leur pertinence clinique et nous avons utilisé des poids pour calculer une mesure de similarité

pondérée entre les patientes. Nous avons également fait varier manuellement les poids des variables pour affiner les résultats de la deuxième méthode à l'aide d'une interface de visualisation interactive de type "Bokeh". Cela nous a permis d'améliorer la compréhension de la similarité entre les patientes.

En utilisant ces méthodes, nous avons construit une matrice de similarité entre les patientes et utilisé du clustering hiérarchique pour les regrouper.

4. **Apprentissage de similarité et optimisation** Suite à la mesure générique, notre objectif était d'améliorer les résultats des 2 méthodes en affinant les mesures de similarité. Nous avons utilisé 2 méthodes différentes à cette fin :

   - *Méthode 1, mesure unique* : Pour la mesure unique, nous avons exploité les réseaux siamois, qui sont des architectures de réseaux neuronaux pour le calcul de similarité. L'idée est d'apprendre les « embeddings » (intégrations) des attributs des patientes. Le processus d'entraînement impliquait des mises à jour itératives, en ajoutant un patient à la fois, et en calculant la fonction de perte (triplet). L'objectif était de créer des « embeddings » qui représentaient efficacement les attributs des patients en utilisant la distance Euclidienne.
   - *Méthode 2, mesure hybride* : Dans cette méthode, des mesures de similarité spécifiques ont été utilisées en fonction des types de variables, rendant l'optimisation par réseau neuronal impraticable. Nous avons opté pour une méthode d'optimisation basée sur Optuna (Akiba *et al.*, 2019) pour automatiser le calcul des poids. L'objectif était de maximiser l'accord avec les annotations d'experts, mesuré à l'aide de l'Indice de Rand Ajusté (ARI).

5. **Evaluation** : Après avoir obtenu des matrices de similarité à l'aide de diverses méthodes, nous disposions d'un total de 7 matrices de similarité, chacune remplissant une fonction différente :

   - $M1^{\text{Euclidean}}$, $M1^{\text{Cosine}}$, $M1^{\text{Jaccard}}$: représentent les matrices de similarité obtenues en utilisant respectivement la similarité euclidienne, la similarité Cosinusienne et la similarité de Jaccard comme mesure unique de calcul.
   - $M1^{\text{Siamese network}}$: la matrice de similarité est obtenue selon la méthode à base de réseaux siamois ;
   - $M2^{\text{NoWeights}}$: la matrice de similarité est obtenue en calculant la moyenne pondérée des trois mesures selon la méthode 2 (mesure hybride), avec une pondération de 1.
   - $M2^{\text{Weighted manually}}$: la matrice de similarité est obtenue selon la méthode 2, avec des poids obtenus manuellement.
   - $M2^{\text{Weighted automatically}}$: la matrice de similarité est obtenue en suivant la méthode 2, avec des poids obtenus en utilisant l'optimisation Optuna .

   Pour l'évaluation, nous avons utilisé le jeu de données « top 5 gold standard ». Nous avons comparé les 5 patientes annotées comme similaires par l'expert (le « gold standard ») avec les 5 patientes les plus similaires fournies par les méthodes de similarité. La métrique utilisée pour l'évaluation était la précision.

Résultats

La Figure 6.5 présente les performances des diverses méthodes sur l'ensemble de test, en utilisant la précision comme mesure d'intérêt. Dans la méthode de mesure unique, les scores de précision variaient de 20 % pour les méthodes Euclidienne et Cosine à presque 30 % pour Jaccard, indiquant ainsi la nécessité d'améliorations.

La Méthode 2 visait à évaluer une approche hybride avec et sans poids. Cette approche a montré des progrès, atteignant un score de précision de 30 % sans poids et 53 % avec des calculs de poids manuels.

L'optimisation a été testée pour maximiser le potentiel de chaque méthode. Alors que l'usage du Réseau Siamois n'a pas amélioré l'utilisation de la distance Euclidienne, l'algorithme d'Optuna pour la Méthode 2 a considérablement amélioré le score de précision, surpassant les autres méthodes et atteignant 68 %.

Discussion

Dans ce travail, notre objectif était d'implémenter une méthode efficace pour la détection de patients similaires à un patient donné. Deux méthodes ont été utilisées : une mesure de similarité unique et une moyenne pondérée de différentes mesures. Les résultats suggèrent que l'adaptation des mesures de similarité aux types de variables et l'application de moyennes pondérées peuvent améliorer les résultats et fournir des informations sur l'importance des variables.

Les résultats du Réseau Siamois étaient suboptimaux, soulignant la nécessité d'un affinage ultérieur et l'augmentation du corpus d'entrainement. Cette approche conserve encore un potentiel pour des performances améliorées.

Enfin, même si nous avons obtenu des chiffres pour la précision, il est essentiel de confirmer que la méthode fonctionne réellement. Pour ce faire, nous devons montrer des cas de patientes similaires obtenus par les algorithmes à des experts et leur demander si l'algorithme propose des profils qui sont vraiment similaires ou pas. De plus, il est essentiel d'augmenter le nombre de F-RCP utilisées pour l'évaluation. Dans cette étude, le processus d'annotation impliquait la sélection des 5 patients les plus similaires, ce qui s'est avéré très chronophage. Il a fallu qu'un expert compare chaque F-RCP de l'ensemble du corpus d'entrainement. Par conséquent, nous avons été limités dans notre capacité à annoter davantage de F-RCP pour la validation. En outre, l'implication de plusieurs experts dans le processus d'annotation pourrait potentiellement améliorer nos résultats.

Conclusion

En conclusion, cette recherche met en lumière l'importance du choix de mesures de similarité appropriées et de méthodes de pondération adaptées aux types des variables. Cette approche améliore la précision de l'identification des patientes similaires tout en fournissant des explications pour les résultats. L'intégration des connaissances d'experts dans le processus de pondération améliore les scores de précision et facilite l'interprétation de la pertinence clinique.

6. Synthèse

L'objectif de cette recherche a été de créer un outil informatisé multifacette d'aide à la décision pour les cliniciens des RCP pour la prise en charge des patients atteints de cancer du sein. Nous

avons développé un SADM adaptatif à la compléxité des cas à traiter, utilisant distinctement des approches basée sur les GBP pour les cas non complexes et basée les cas similaires pour traiter les cas complexes.

Tout au long de la thèse, nous avons exploré des aspects clés de l'aide à la décision clinique, tels que l'extraction de données à partir de documents cliniques non structurés, la classification de la complexité des cas et les différentes méthodes d'aide à la décision.

- Nous avons mis au point une méthode basée sur des règles pour extraire des données structurées des fiches RCP. Le processus de traitement a montré de bonnes performances, mettant en lumière des attributs pour lesquels les méthodes basées sur les règles peuvent être limitées, suggérant l'utilisation de méthodes d'apprentissage approfondies pour ces attributs (voir annexe A) (Chapitre 3).

- Lors de la classification de la complexité des cas, les modèles d'apprentissage automatique traditionnels ont surpassé les modèles de transformers dans ce cas d'usage, soutenant ainsi l'approche basée sur l'utilisation d'annotateurs sémantiques pour l'extraction de caractéristiques (Chapitre 4).

- Pour les cas non complexes, le SADM basé sur le GBP du projet DESIREE a été mis à jour en utilisant de vrais cas de cancer du sein. Nous avons proposé une approche novatrice permettant d'identifier efficacement les profils les plus courants, facilitant ainsi la mise à jour nécessaire de la base de connaissances du GL-DSS (Chapitre 5).

- Pour les cas complexes, La méthode basée sur une mesure de similarité hybride a montré des résultats prometteurs dans la détection de patients similaires. Cependant, des améliorations sont nécessaires, notamment la vérification de la pertinence des résultats par des experts (Chapitre 6).

Malgré ces avancées, notre étude présente des limites liées à la taille de l'échantillon, à la généralisation, à la qualité des données et aux difficultés techniques. L'élargissement de l'ensemble de données et de la portée de la recherche pourrait améliorer la généralisation de nos résultats. Les contraintes pratiques et temporelles ont limité l'implication approfondie des experts.

Pour l'avenir, des perspectives de recherche prometteuses incluent l'amélioration de la qualité des données, la validation des systèmes d'aide à la décision avec des retours externes, l'optimisation des méthodes basées sur la similarité et l'extension à d'autres types de cancer.
L'intégration de modèles de langage de grande taille (LLMs) pourrait automatiser la mise à jour des GBP et simplifier le processus d'aide à la décision.

Bibliography

AAMODT A. & PLAZA E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, **7**(1), 39–59. 25, 161

ABIDI S. R. (2011). Ontology-based knowledge modeling to provide decision support for comorbid diseases. In D. RIAÑO, A. TEN TEIJE, S. MIKSCH & M. PELEG, Coordinateurs, *Knowledge Representation for Health-Care*, p. 27–39, Berlin, Heidelberg: Springer Berlin Heidelberg. 19

ABIDI S. R., ABIDI S. S. R., HUSSAIN S. & SHEPHERD M. (2007). Ontology-based modeling of clinical practice guidelines: a clinical decision support system for breast cancer follow-up interventions at primary care settings. *Studies in Health Technology and Informatics*, **129**(Pt 2), 845–849. 19

ADEBOYEJE G., AGIRO A., MALIN J., FISCH M. J. & DEVRIES A. (2017). Reducing Overuse of Colony-Stimulating Factors in Patients With Lung Cancer Receiving Chemotherapy: Evidence From a Decision Support-Enabled Program. *Journal of Oncology Practice*, **13**(4), e337–e345. 13

ADEVA J. G., ATXA J. P., CARRILLO M. U. & ZENGOTITABENGOA E. A. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, **41**, 1498–1508. 35

AGHAREZAEI Z., TOFIGHI S., NEMATI A., AGHAREZAEI L. & BAHAADINBEIGI K. (2013). Surveying kerman's afazalipour hospital clinical and educational staff's points of view about the clinical decision support system designed for reducing the possibility of pulmonary embolism and deep vein thrombosis. *Hospital Quarterly*, **12**(2), 29–38. 12

AHMED M. U., BEGUM S., OLSSON E., XIONG N. & FUNK P. (2010). *Case-Based Reasoning for Medical and Industrial Decision Support Systems*, In *Successful Case-based Reasoning Applications-I*, p. 7–52. 25

AKBIK A., BLYTHE D. & VOLLGRAF R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1638–1649. 29

AKIBA T., SANO S., YANASE T., OHTA T. & KOYAMA M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. Issue: arXiv:1907.10902 arXiv: 1907.10902 [cs, stat]. 98, 99, 131

ALAWAD M., YOON H. & TOURASSI G. (2018). Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports. In *2018 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2018*, p. 218–221. 32

ALFONSECA E. & MANANDHAR S. (2002). An unsupervised method for general named entity recognition and automated concept discovery. In *Proceedings of the 1st International Conference on General WordNet Mysore India*, volume 69(6), p. 1–9. 30

ALGHOSON A. M. (2014). Medical Document Classification Based on MeSH. In *Proceedings of the 2014 forty-seventh Hawaii international conference on system sciences*, p. 2571–2575: IEEE. 36

ALSENTZER E., MURPHY J., BOAG W., WENG W.-H., JINDI D., NAUMANN T. & MCDERMOTT M. (2019). Publicly Available Clinical. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, p. 72–78. 30

AOUINA O., HILBEY J. & CHARLET J. (2023). Ontology-Based Semantic Annotation of French Psychiatric Clinical Documents. *Studies in Health Technology and Informatics*, **302**, 793–797. 33

ARBOUR K. C., LUU A. T., LUO J., RIZVI H., PLODKOWSKI A. J., SAKHI M., HUANG K. B., DIGUMARTHY S. R., GINSBERG M. S., GIRSHMAN J., KRIS M. G., RIELY G. J., YALA A., GAINOR J. F., BARZILAY R. & HELLMANN M. D. (2021). Deep Learning to Estimate RECIST in Patients with NSCLC Treated with PD-1 Blockade. *Cancer Discovery*, **11**(1), 59–67. 14

ARMENGOL E., ESTEVA F., GODO L. & TORRA (2004). On learning similarity relations in fuzzy case-based reasoning. *Transactions on Rough Sets II, Lecture Notes in Computer Science*, **3135**, 14–32. 25

ARNOLD M., MORGAN E., RUMGAY H., MAFRA A., SINGH D., LAVERSANNE M., VIGNAT J., GRALOW J. R., CARDOSO F., SIESLING S. & SOERJOMATARAM I. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. *The Breast*, **66**, 15–23. 1, 113

ARONSON A. & LANG F.-M. (2010). An overview of metamap: historical perspective and recent advances. *JAMIA*, **17**(3), 229–236. 32, 64, 65

ARONSON A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, p.Ĩ7: American Medical Informatics Association. 65

AZARPIRA M., REDJDAL A., BOUAUD J. & SEROUSSI B. (2022). Methods Used to Compare Narrative Clinical Practice Guidelines: A Scoping Review. *Studies in Health Technology and Informatics*, **295**, 304–307. 8, 20, 76

BATES J., FODEH S. J., BRANDT C. A. & WOMACK J. A. (2015). Classification of radiology reports for falls in an HIV study cohort. *Journal of the American Medical Informatics Association*, **23**, e113–e117. 34, 36

BEAUCHEMIN M., MURRAY M. T., SUNG L. & ET AL. (2019). Clinical decision support for therapeutic decision-making in cancer: A systematic review. *Int J Med Inform*, **130**, 103940. 3, 114

BELTAGY I., LO K. & COHAN A. (2020a). Scibert: A pretrained language model for scientific text. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, p. 3615–3620. 29

BELTAGY I., PETERS M. & COHAN A. (2020b). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*. 72

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & SHMITCHELL S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, p. 610–623. 29

BENESTY J., CHEN J., HUANG Y. & COHEN I. (2009). Pearson correlation coefficient. In *Noise Reduction in Speech Processing: 2*, p. 1–4. Springer. 36

BERTSCHE T., ASKOXYLAKIS V., HABL G., LAIDIG F., KALTSCHMIDT J., SCHMITT S. P. W., GHADERI H., BOIS A. Z.-D., MILKER-ZABEL S., DEBUS J., BARDENHEUER H. J. & HAEFELI W. E. (2009). Multidisciplinary pain management based on a computerized clinical decision support system in cancer pain patients. *PAIN®*, **147**(1), 20–28. 13

BIBAULT J.-E., BASSENNE M., REN H. & XING L. (2020). Deep Learning Prediction of Cancer Prevalence from Satellite Imagery. *Cancers*, **12**(12), 3844. 14

BILICI E., DESPOTOU G. & ARVANITIS T. N. (2018). The use of computer-interpretable clinical guidelines to manage care complexities of patients with multimorbid conditions: A review. *DIGITAL HEALTH*, **4**, 2055207618804927. Publisher: SAGE Publications Ltd. 13, 17

BITTERMAN D. S., MILLER T. A., MAK R. H. & SAVOVA G. K. (2021). Clinical natural language processing for radiation oncology: A review and practical primer. *International Journal of Radiation Oncology Biology Physics*, **110(3)**, 641–655. 32

BLAYNEY D. W. (2013). Tumor boards (team huddles) aren't enough to reach the goal. *Journal of the National Cancer Institute*, **105**(2), 82–84. 2, 113

BODENREIDER O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, **32**(DATABASE ISS.), D267. 29

BODENREIDER O. (2008). Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. *Yearbook of medical informatics*, p. 67–79. 19

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *TACL*, **5**, 135–146. 29, 64

BONATTI E., KUCHUKHIDZE G. & ZAMARIAN L. (2009). Decision making in ambiguous and risky situations after unilateral temporal lobe epilepsy surgery. *Epilepsy & Behavior*, **14**, 665–673. 12

BOUAUD J., MESSAI N., LAOUÉNAN C. & AL E. (2012). Elicitating patient patterns of physician noncompliance with breast cancer guidelines using formal concept analysis. *Stud Health Technol Inform*, **180**, 477–481. 13

BOUAUD J., PELAYO S., LAMY J.-B. & ET AL. (2020a). Implementation of an ontological reasoning to support the guideline-based management of primary breast cancer patients in the desiree project. *Artif Intell Med*, **108**, 101922. 5, 116

BOUAUD J., PELAYO S., LAMY J.-B., PREBET C., NGO C., TEIXEIRA L., GUÉZENNEC G. & SEROUSSI B. (2020b). Implementation of an ontological reasoning to support the guideline-based management of primary breast cancer patients in the DESIREE project. *Artificial Intelligence in Medicine*, **108**, 101922. 3, 16, 20, 21, 22, 43, 115, 120, 161

BOUAUD J., S&#233, ROUSSI B., BRIZON A., CULTY T., MENTR&#233, FRANCE & RAVERY V. (2007). How Updating Textual Clinical Practice Guidelines Impacts Clinical Decision Support Systems: a Case Study with Bladder Cancer Management. In *MEDINFO 2007*, p. 829–833. IOS Press. 76, 77, 162

BRAR S. S., HONG N. L. & WRIGHT F. C. (2014). Multidisciplinary cancer care: does it improve outcomes? *Journal of Surgical Oncology*, **110**(5), 494–499. 2, 113

BRIN S. (1999). Extracting patterns and relations from the world wide web. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 1590, p. 172–183. 30

BROWN S.-A. (2016). Patient Similarity: Emerging Concepts in Systems and Precision Medicine. *Frontiers in Physiology*, **7**, 561. 104, 129

BROWN T. B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020a). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*. 29

BROWN T. B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. & OTHERS (2020b). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*. 30

BUCHAN K., FILANNINO M. & UZUNER O. (2017). Automatic prediction of coronary artery disease from clinical narratives. In *Journal of Biomedical Informatics*, volume 72, p. 23–32. 35

BURGER G., ABU-HANNA A., KEIZER N. D. & CORNET R. (2016). Natural language processing in pathology: a scoping review. *Journal of Clinical Pathology*, **69**(11), 949–955. Publisher: BMJ Publishing Group Section: Review. 32

BUTT L., ZUCCON G., NGUYEN A., BERGHEIM A. & GRAYSON N. (2013). Classification of cancer-related death certificates using machine learning. *Australasian Medical Journal*, **6**, 292–300. 34, 36

CAMPBELL H. E., TAYLOR M. A., HARRIS A. L. & GRAY A. M. (2009). An investigation into the performance of the Adjuvant! Online prognostic programme in early breast cancer for a cohort of patients in the United Kingdom. *British Journal of Cancer*, **101**(7), 1074–1084. 15

CANDIDO DOS REIS F. J., WISHART G. C., DICKS E. M., GREENBERG D., RASHBASS J., SCHMIDT M. K., VAN DEN BROEK A. J., ELLIS I. O., GREEN A., RAKHA E., MAISHMAN T., ECCLES D. M. & PHAROAH P. D. P. (2017). An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast cancer research: BCR*, **19**(1), 58. 15

CANO C., MONAGHAN T., BLANCO A., WALL D. P. & PESHKIN L. (2009). Collaborative text-annotation resource for disease-centered relation extraction from biomedical text. *Journal of Biomedical Informatics*, **42**(5), 967–977. 31

CEJUELA J. M., MCQUILTON P., PONTING L., MARYGOLD S. J., STEFANCSIK R., MILLBURN G. H., ROST B. & THE FLYBASE CONSORTIUM (2014). tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database*, **2014**, bau033. 31

CHANG Y., LAFATA K., SUN W., WANG C., CHANG Z., KIRKPATRICK J. P. & YIN F.-F. (2019). An investigation of machine learning methods in delta-radiomics feature analysis. *PloS One*, **14**(12), e0226348. 14

CHAPMAN W. W., BRIDEWELL W., HANBURY P., COOPER G. F. & BUCHANAN B. G. (2001). A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, **34**(5), 301–310. 51

CHEN T. & GUESTRIN C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: ACM. 66

CHEN X., FAVIEZ C., VINCENT M., BRISEÑO-ROA L., FAOUR H., ANNEREAU J.-P., LYONNET S., ZAIDAN M., SAUNIER S., GARCELON N. & BURGUN A. (2022). Patient-Patient Similarity-Based Screening of a Clinical Data Warehouse to Support Ciliopathy Diagnosis. *Frontiers in Pharmacology*, **13**, 786710. 26

CHOUDHURY N. & ARA S. (2016). A Survey on Case-based Reasoning in Medicine. *International Journal of Advanced Computer Science and Applications*, **7**(8). 25, 27

CIPRUT S. E., KELLY M. D., WALTER D., HOFFMAN R., BECKER D. J., LOEB S., SEDLANDER E., TENNER C. T., SHERMAN S. E., ZELIADT S. B. & MAKAROV D. V. (2020). A Clinical Reminder Order Check Intervention to Improve Guideline-concordant Imaging Practices for Men With Prostate Cancer: A Pilot Study. *Urology*, **145**, 113–119. 13

CLARK C., WELLNER B., DAVIS R., ABERDEEN J. & HIRSCHMAN L. (2017). Automatic classification of RDoC positive valence severity with a neural network. *Journal of Biomedical Informatics*. 35

CLARK K., LUONG M.-T., LE Q. V. & MANNING C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*. 29

CODEN A., SAVOVA G., SOMINSKY I., TANENBLATT M., MASANZ J., SCHULER K., COOPER J., GUAN W. & DE GROEN P. C. (2009). Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *Journal of Biomedical Informatics*, **42**(5), 937–949. 40

COHEN R., AVIRAM I., ELHADAD M. & ELHADAD N. (2014). Redundancy-aware topic modeling for patient record notes. *PloS One*, **9**(2), e87555. 63

COLLINS M. & SINGER Y. (1999). Unsupervised models for named entity classification. In *Proceedings of EMNLP/VLC-99*. 30

COLLOBERT R. & WESTON J. (2008). A unified architecture for natural language processing. In *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*, p. 160–167: ACM Press. 29

COLLOBERT R., WESTON J., BOTTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, **12**, 2493–2537. 29

CORTI C., COBANAJ M., DEE E. C., CRISCITIELLO C., TOLANEY S. M., CELI L. A. & CURIGLIANO G. (2023). Artificial intelligence in cancer research and precision medicine: Applications, limitations and priorities to drive transformation in the delivery of equitable and unbiased care. *Cancer Treatment Reviews*, **112**, 102498. 3

CUNNINGHAM H., MAYNARD D., BONTCHEVA K. & TABLAN V. (2002). *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. Journal Abbreviation: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02) Publication Title: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). 31

CÁNOVAS-SEGURA B., MORALES A., JUAREZ J. M., CAMPOS M. & PALACIOS F. (2019). A lightweight acquisition of expert rules for interoperable clinical decision support systems. *Knowledge-Based Systems*, **167**, 98–113. 13, 76

DALLOUX C., CLAVEAU V. & GRABAR N. (2017). Détection de la négation : corpus français et apprentissage supervisé. p.1̃. 51

DANIYAL A., ABIDI S. R. & ABIDI S. S. R. (2009). Computerizing clinical pathways: ontology-based modeling and execution. *Studies in Health Technology and Informatics*, **150**, 643–647. 19

DANSO S., ATWELL E. & JOHNSON O. (2013). Linguistic and statistically derived features for cause of death prediction from verbal autopsy text. In *Language processing and knowledge in the web*, p. 47–60: Springer. 34, 35

DATTA S., BERNSTAM E. V. & ROBERTS K. (2019). A frame semantic overview of nlp-based information extraction for cancer-related ehr notes. 32

DATTA S., SI Y., RODRIGUEZ L., SHOOSHAN S. E., DEMNER-FUSHMAN D. & ROBERTS K. (2020). Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest x-ray reports using deep learning. *Journal of Biomedical Informatics*, **108**(February), 103473. 32

DE CAO N., IZACARD G., RIEDEL S. & PETRONI F. (2020). Autoregressive entity retrieval. In *9th International Conference on Learning Representations*. 30

DE CLERCQ P. A., BLOM J. A., KORSTEN H. H. & HASMAN A. (2004). Approaches for creating computer-interpretable guidelines that facilitate decision support. *Artificial Intelligence in Medicine*, **31**(1), 1–27. 18

DE MANTARAS R. L. (2001). *Case-based reasoning*. Springer Berlin Heidelberg. 24, 92

DE MANTARAS R. L. *et al.* (2005). Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review*, **20**(03), 215–240. 25

DEBOLE F. & SEBASTIANI F. (2004). Supervised term weighting for automated text categorization. In *Text mining and its applications*, p. 81–97: Springer. 35

DENG Y., GROLL M. J. & DENECKE K. (2015). Rule-based cervical spine defect classification using medical narratives. *Studies in health technology and informatics*, **216**, 1038. 36

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, **1**, 4171–4186. 29, 64

DIGAN W., WACK M., LOOTEN V., NEURAZ A., BURGUN A. & RANCE B. (2019). Evaluating the Impact of Text Duplications on a Corpus of More than 600,000 Clinical Narratives in a French Hospital. *Studies in Health Technology and Informatics*, **264**, 103–107. 33

DONNELLY K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, **121**, 279. Publisher: IOS Press. 35

DOSHI-VELEZ F. & KIM B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [cs, stat]. 14

DOWSETT M., SESTAK I., REGAN M. M., DODSON A., VIALE G., THÜRLIMANN B., COLLEONI M. & CUZICK J. (2018). Integration of Clinical Variables for the Prediction of Late Distant Recurrence in Patients With Estrogen Receptor-Positive Breast Cancer Treated With 5 Years of Endocrine Therapy: CTS5. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, **36**(19), 1941–1948. 15

DURA B., DURKIN K., GUEDJ R., GOTTLIEB D., GUNNING D., KIM E., KIM J., KIM J., MOHTA K., REDDY V., RUDIN C., SO D. & SONTAG D. (2022a). Learning structures of the french clinical language: development and validation of word embedding models using 21 million clinical reports from electronic health records. *arXiv*. 67

DURA B., WAJSBURT P., PETIT-JEAN T., COHEN A., JEAN C. & BEY R. (2022b). Eds-nlp: efficient information extraction from french clinical notes. 34, 49, 122

ECCHER C., SEYFANG A. & FERRO A. (2014). Implementation and evaluation of an Asbru-based decision support system for adjuvant treatment in breast cancer. *Computer Methods and Programs in Biomedicine*, **117**(2), 308–321. 15

EFTIMOV T., KOROUŠIĆ SELJAK B. & KOROŠEC P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS ONE*, **12**(6), e0179488. 20

EL SAGHIR N. S., ASSI H. A., KHOURY K. E., EL ZAWAWY A. M., ABBAS J. A. & EID T. A. (2013). Re: Tumor boards and the quality of cancer care. *Journal of the National Cancer Institute*, **105**(23), 1839. 2, 113

ETZIONI O., CAFARELLA M., DOWNEY D., POPESCU A.-M. M., SHAKED T., SODERLAND S., WELD D. & YATES A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, **165**(1), 91–134. 30

EVIDENCE-BASED MEDICINE WORKING GROUP (1992). Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA*, **268**(17), 2420–2425. 2, 114

FERNANDES M., VIEIRA S. M., LEITE F., PALOS C., FINKELSTEIN S. & SOUSA J. M. C. (2020). Clinical Decision Support Systems for Triage in the Emergency Department using Intelligent Systems: a Review. *Artificial Intelligence in Medicine*, **102**, 101762. 11

FINKEL J. & MANNING C. (2009). Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 1 - EMNLP '09*, volume 1, p. 141, Morristown, NJ, USA: Association for Computational Linguistics. 30

FINNIE G. & SUN Z. (2003). A logical foundation for the case-based reasoning cycle. *International journal of intelligent systems*, **18**(4), 367–382. 25

FIRTH J. (1957). A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, **1952-59**, 1–32. 29

FORMAN G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, **3**, 1289–1305. 36

FORSYTH A. W., BARZILAY R., HUGHES K. S., LUI D., LORENZ K. A., ENZINGER A., TULSKY J. A. & LINDVALL C. (2018). Machine Learning Methods to Extract Documentation of Breast Cancer Symptoms From Electronic Health Records. *Journal of Pain and Symptom Management*, **55**(6), 1492–1499. 40

FOX J., JOHNS N., LYONS C., RAHMANZADEH A., THOMSON R. & WILSON P. (1997). PROforma: a general technology for clinical decision support systems. *Computer Methods and Programs in Biomedicine*, **54**(1), 59–67. 18

FRAILE NAVARRO D., IJAZ K., REZAZADEGAN D., RAHIMI-ARDABILI H., DRAS M., COIERA E. & BERKOVSKY S. (2023). Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, **177**, 105122. 30

FUX A., SOFFER P. & PELEG M. (2020). A layered computer-interpretable guideline model for easing the update of locally adapted clinical guidelines. *Health Informatics Journal*, **26**(1), 156–171. 13

GALOPIN A. (2015). *Modélisation ontologique des recommandations de pratique clinique pour une aide à la décision à niveaux d'abstraction variables*. phdthesis, Université Pierre et Marie Curie - Paris VI. 19

GALOPIN A., BOUAUD J., PEREIRA S. & SÉROUSSI B. (2014a). Comparison of clinical practice guidelines from a knowledge modelling perspective: a case study with the management of hypertension. *Stud Health Technol Inform*, **197**, 21–25. 20, 21

GALOPIN A., BOUAUD J., PEREIRA S. & SÉROUSSI B. (2014b). Using an ontological modeling to evaluate the consistency of clinical practice guidelines: application to the comparison of three guidelines on the management of adult hypertension. *Stud Health Technol Inform*, **205**, 38–42. 20, 21

GALOPIN A., BOUAUD J., PEREIRA S. & SEROUSSI B. (2015). An ontology-based clinical decision support system for the management of patients with multiple chronic disorders. *Stud Health Technol Inform*, **216**, 275–279. 19, 20, 21

GAMEIRO G. R., SINKUNAS V., LIGUORI G. R. & AULER-JÚNIOR J. O. C. (2018). Precision Medicine: Changing the way we think about healthcare. *Clinics*, **73**, e723. 2, 114

GAO S., ALAWAD M., YOUNG M., GOUNLEY J., SCHAEFFERKOETTER N., YOON H., WU X., DURBIN E., DOHERTY J., STROUP A. & COYLE L. (2021a). Limitations of transformers on clinical text classification. *IEEE journal of biomedical and health informatics*, **25**(9), 3596–3607. 20, 72

GAO S., ALAWAD M., YOUNG M., GOUNLEY J., SCHAEFFERKOETTER N., YOON H., WU X., DURBIN E., DOHERTY J., STROUP A., COYLE L. & TOURASSI G. (2021b). Limitations of transformers on clinical text classification. *IEEE J Biomed Health Inform*, **25**(9), 3596–3607. 67

GEORGIA_RESEARCH_INSTITUTE (2018). ClarityNLP. https://github.com/ClarityNLP/ClarityNLP. 50, 122

GOLDBERG Y. & LEVY O. (2014). word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*. 35, 67

GOOCH P. & ROUDSARI A. (2011). Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems. *J Am Med Inform Assoc*, **18**, 738–748. 16

GORDON C. (1996). May we support your decision? *Journal of Health Services Research & Policy*, **1**(3), 175–178. 13

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS: French Corpus with Clinical Cases. p.Ĩ. 51

GREENES R. (2014). *Clinical Decision Support: The Road to Broad Adoption: Second Edition*. Pages: 887. 12

GROUIN C., DELÉGER L., ROSIER A., TEMAL L., DAMERON O., VAN HILLE P., BURGUN A. & ZWEIGENBAUM P. (2011). Automatic computation of CHA2DS2-VASc score: information extraction from clinical texts for thromboembolism risk assessment. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, **2011**, 501–510. 33

GROSSWENDT A., RÖGLIN H. & SCHMIDT M. (2019). Analysis of Ward's Method. Issue: arXiv:1907.05094 arXiv: 1907.05094 [cs]. 97

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J., POON H. & OTHERS (2020). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *arXiv preprint arXiv:2007.15779*. 30

GU Y., YANG X., TIAN L., YANG H., LV J., YANG C., WANG J., XI J., KONG G. & ZHANG W. (2022). Structure-aware siamese graph neural networks for encounter-level patient similarity learning. *Journal of Biomedical Informatics*, **127**, 104027. 104, 105

GUARINO N., OBERLE D. & STAAB S. (2009). What Is an Ontology? In *Handbook on Ontologies*, p. 1–17. Journal Abbreviation: Handbook on Ontologies. 19

GUYON I. & ELISSEEFF A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, **3**, 1157–1182. 36

GÉRARDIN C., MAGEAU A., MÉKINIAN A., TANNIER X. & CARRAT F. (2022a). Construction of Cohorts of Similar Patients From Automatic Extraction of Medical Concepts: Phenotype Extraction Study. *JMIR Medical Informatics*, **10**(12), e42379. 27

GÉRARDIN C., WAJSBÜRT P., VAILLANT P., BELLAMINE A., CARRAT F. & TANNIER X. (2022b). Multilabel classification of medical concepts for patient clinical profile identification. *Artificial Intelligence in Medicine*, **128**, 102311. 33

HAINES A. & JONES R. (1994). Implementing findings of research. *BMJ (Clinical research ed.)*, **308**(6942), 1488–1492. 13

HAMMOND P., HARRIS A. L., DAS S. K. & WYATT J. C. (1994). Safety and decision support in oncology. *Methods of Information in Medicine*, **33**(4), 371–381. 2

HARRIS Z. (1954). Distributional structure. *WORD*, **10**(3), 146–162. 29

HE T., PUPPALA M., OGUNTI R., MANCUSO J., YU X., CHEN S., CHANG J., PATEL T. & WONG S. (2017). Deep learning analytics for diagnostic support of breast cancer disease management. In *2017 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2017*, p. 365–368. 32

HENDRIKS M. P., VERBEEK X. A. A. M., VAN MANEN J. G., VAN DER HEIJDEN S. E., GO S. H. L., GOOIKER G. A., VAN VEGCHEL T., SIESLING S. & JAGER A. (2020). Clinical decision trees support systematic evaluation of multidisciplinary team recommendations. *Breast Cancer Research and Treatment*, **183**(2), 355–363. 15

HENDRIKS M. P., VERBEEK X. A. A. M., VAN VEGCHEL T., VAN DER SANGEN M. J. C., STROBBE L. J. A., MERKUS J. W. S., ZONDERLAND H. M., SMORENBURG C. H., JAGER A. & SIESLING S. (2019). Transformation of the National Breast Cancer Guideline Into Data-Driven Clinical Decision Trees. *JCO clinical cancer informatics*, **3**, 1–14. 15

HOFFER E. & AILON N. (2018). Deep metric learning using Triplet network. arXiv:1412.6622 [cs, stat]. 28

HOLT A., BICHINDARITZ I., SCHMIDT R. & PERNER P. (2005). Medical applications in case-based reasoning. *The Knowledge Engineering Review*, **20**(03), 289–292. 25

HONG J. C., ECLOV N. C. W., DALAL N. H., THOMAS S. M., STEPHENS S. J., MALICKI M., SHIELDS S., COBB A., MOWERY Y. M., NIEDZWIECKI D., TENENBAUM J. D. & PALTA M. (2020). System for High-Intensity Evaluation During Radiation Therapy (SHIELD-RT): A Prospective Randomized Study of Machine Learning-Directed Clinical Evaluations During Radiation and Chemoradiation. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, **38**(31), 3652–3661. 14

HUANG H.-Z., LU X.-D., GUO W., JIANG X.-B., YAN Z.-M. & WANG S.-P. (2021). Heterogeneous Information Network-Based Patient Similarity Search. *Frontiers in Cell and Developmental Biology*, **9**, 735687. 26

HUANG Z., XU W., KAI Y. & YU K. (2015). Bidirectional lstm-crf models for sequence tagging. *CoRR*, **abs/1508.0**. 30

HUSSAIN M. & LEE S. (2019). Information extraction from clinical practice guidelines: A step towards guidelines adherence. In *Proceedings of the 2019*, p. 1029–1036. 20

JACOB C., SANCHEZ-VAZQUEZ A. & IVORY C. (2019). Clinicians' Role in the Adoption of an Oncology Decision Support App in Europe and Its Implications for Organizational Practices: Qualitative Case Study. *JMIR mHealth and uHealth*, **7**(5), e13555. 15

JENDERS R. A., CORMAN R. & DASGUPTA B. (2003). Making the Standard More Standard: A Data and Query Model for Knowledge Representation in the Arden Syntax. *AMIA Annual Symposium Proceedings*, **2003**, 323–327. 18

JENSEN K., SOGUERO-RUIZ C., OYVIND MIKALSEN K. & AL E. (2017). Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci Rep*, **7**, 46226. 16

JIANG X., WELLS A., BRUFSKY A. & NEAPOLITAN R. (2019). A clinical decision support system learned from data to personalize treatment recommendations towards preventing breast cancer metastasis. *PloS One*, **14**(3), e0213292. 17

JIE Z., ZHIYING Z. & LI L. (2021). A meta-analysis of Watson for Oncology in clinical application. *Scientific Reports*, **11**(1), 5792. 15

JO T. (2013). Application of table based similarity to classification of bio-medical documents. In *Proceedings of the 2013 IEEE International Conference on Granular Computing (GRC)*, p. 162–166: IEEE. 35

JOUFFROY J., FELDMAN S. F., LERNER I., RANCE B., BURGUN A. & NEURAZ A. (2021). Hybrid Deep Learning for Medication-Related Information Extraction From Clinical Texts in French: MedExt Algorithm Development Study. *JMIR medical informatics*, **9**(3), e17934. 33

KALTER H. D., PERIN J. & BLACK R. E. (2016). Validating hierarchical verbal autopsy expert algorithms in a large data set with known causes of death. *Journal of Global Health*, **6**(1), 010601. Publisher: Edinburgh University Global Health Society. 36

KANG J., MORIN O. & HONG J. C. (2020). Closing the Gap Between Machine Learning and Clinical Cancer Care-First Steps Into a Larger World. *JAMA oncology*, **6**(11), 1731–1732. 14

KANN B. H., HICKS D. F., PAYABVASH S., MAHAJAN A., DU J., GUPTA V., PARK H. S., YU J. B., YARBROUGH W. G., BURTNESS B. A., HUSAIN Z. A. & ANEJA S. (2020). Multi-Institutional Validation of Deep Learning for Pretreatment Identification of Extranodal Extension in Head and Neck Squamous Cell Carcinoma. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, **38**(12), 1304–1311. 14

KANN B. H., HOSNY A. & AERTS H. J. (2021). Artificial Intelligence for Clinical Oncology. *Cancer cell*, **39**(7), 916–927. 14

KASTHURIRATHNE S. N., DIXON B. E., GICHOYA J., XU H., XIA Y., MAMLIN B. & OTHERS (2016). Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection. *Journal of Biomedical Informatics*, **60**, 145–152. Publisher: Elsevier. 36

KASTHURIRATHNE S. N., DIXON B. E., GICHOYA J., XU H., XIA Y., MAMLIN B. & OTHERS (2017). Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data. *Journal of Biomedical Informatics*, **69**, 160–176. Publisher: Elsevier. 34

KATIYAR A. & CARDIE C. (2018). Nested named entity recognition revisited. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, p. 861–871, Stroudsburg, PA, USA: Association for Computational Linguistics. 30

KAVULURU R., RIOS A. & LU Y. (2015). An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial Intelligence in Medicine*, **65**(2), 155–166. Publisher: Elsevier. 35

KEATING N. L., LANDRUM M. B., LAMONT E. B., BOZEMAN S. R., SHULMAN L. N. & MCNEIL B. J. (2013). Tumor boards and the quality of cancer care. *Journal of the National Cancer Institute*, **105**(2), 113–121. 2

KENT D. L., SHORTLIFFE E. H., CARLSON R. W., BISCHOFF M. B. & JACOBS C. D. (1985). Improvements in data collection through physician use of a computer-based chemotherapy treatment consultant. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, **3**(10), 1409–1417. 13

KESSON E. M., ALLARDICE G. M., GEORGE W. D., BURNS H. J. G. & MORRISON D. S. (2012). Effects of multidisciplinary team working on breast cancer survival: retrospective, comparative, interventional cohort study of 13 722 women. *BMJ (Clinical research ed.)*, **344**, e2718. 2, 113

KHAN O., LIM CHOI KEUNG S. N., ZHAO L. & ARVANITIS T. N. (2014). A Hybrid EAV-Relational Model for Consistent and Scalable Capture of Clinical Research Data. *Studies in Health Technology and Informatics*, **202**, 32–35. 43

KICKINGEREDER P., ISENSEE F., TURSUNOVA I., PETERSEN J., NEUBERGER U., BONEKAMP D., BRUGNARA G., SCHELL M., KESSLER T., FOLTYN M., HARTING I., SAHM F., PRAGER M., NOWOSIELSKI M., WICK A., NOLDEN M., RADBRUCH A., DEBUS J., SCHLEMMER H.-P., HEILAND S., PLATTEN M., VON DEIMLING A., VAN DEN BENT M. J., GORLIA T., WICK W., BENDSZUS M. & MAIER-HEIN K. H. (2019). Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *The Lancet. Oncology*, **20**(5), 728–740. 14

KIM D. W., JANG H. Y., KIM K. W., SHIN Y. & PARK S. H. (2019). Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. *Korean Journal of Radiology*, **20**(3), 405–410. 14

KIM J., OHTA T., TATEISI Y. & TSUJII J. (2003). Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, **19**(SUPPL. 1), i180–i182. 30

KING X., LYU I., R.-T. M. & JIN R. (2010). Discriminative semi-supervised feature selection via manifold regularization. *IEEE Transactions on Neural Networks*, **21**, 1033–1047. 36

KLARENBEEK S. E., SCHUURBIERS-SIEBERS O. C. J., VAN DEN HEUVEL M. M. & AL E. (2020a). Barriers and facilitators for implementation of a computerized clinical decision support system in lung cancer multidisciplinary team meetings-a qualitative assessment. *Biology (Basel)*, **10**, 9. 14, 16

KLARENBEEK S. E., WEEKENSTROO H. H. A., SEDELAAR J. P. M., FÜTTERER J. J., PROKOP M. & TUMMERS M. (2020b). The Effect of Higher Level Computerized Clinical Decision Support Systems on Oncology Care: A Systematic Review. *Cancers*, **12**(4), 1032. 13, 17

KLEIN D., SMARR J., NGUYEN H. & MANNING C. (2003). Named entity recognition with character-level models. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* -, volume 4, p. 180–183, Morristown, NJ, USA: Association for Computational Linguistics. 29, 30

KOCBEK S., CAVEDON L., MARTINEZ D., BAIN C., MAC MANUS C., HAFFARI G. & OTHERS (2016). Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources. *Journal of Biomedical Informatics*, **64**, 158–167. Publisher: Elsevier. 35

KONG L., D'AUTUME C., LING W., YU L., DAI Z. & YOGATAMA D. (2020). A mutual information maximization perspective of language representation learning. In *ICLR*. 29

KOOPMAN B., KARIMI S., NGUYEN A., MCGUIRE R., MUSCATELLO D., KEMP M. & OTHERS (2015). Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC medical informatics and decision making*, **15**(1), 10. Publisher: BioMed Central. 35, 36

KOTON P. (1989). A medical reasoning program that improves with experience. *Computer methods and programs in biomedicine*, **30**(2), 177–184. 25

KOUZ H., BOUAUD J., GUÉZENNEC G. & SEROUSSI B. (2020). From Atomic Guideline-Based Recommendations to Complete Therapeutic Care Plans: A Knowledge-Based Approach Applied to Breast Cancer Management. *Studies in Health Technology and Informatics*, **275**, 107–111. 54, 161

KUSNER M. J., SUN Y., KOLKIN N. I. & WEINBERGER K. Q. (2015). From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, p. 957–966: JMLR.org. 27

KWON D., KIM S., SHIN S. & WILBUR J. (2013). BioQRator : a web-based interactive biomedical literature curating system. 31

LACSON R., HARRIS K., BRAWARSKY P., TOSTESON T., ONEGA T., TOSTESON A., KAYE A., GONZALEZ I., BIRDWELL R. & HAAS J. (2015). Evaluation of an automated information extraction tool for imaging data elements to populate a breast cancer screening registry. *Journal of Digital Imaging*, **28**(5), 567. 32

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 260–270, Stroudsburg, PA, USA: Association for Computational Linguistics. 30

LAMY J.-B. (2017). Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artificial Intelligence in Medicine*, **80**, 11–28. 96

LAMY J.-B., SEKAR B., GUEZENNEC G., BOUAUD J. & SÉROUSSI B. (2019). Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*, **94**, 42–53. 27

LANZOLA G., POLCE F., PARIMBELLI E., GABETTA M., CORNET R., DEGROOT R., KOGAN A., GLASSPOOL D., WILK S. & QUAGLINI S. (2023). The Case Manager: An Agent Controlling the Activation of Knowledge Sources in a FHIR-based Distributed Reasoning Environment. *Applied Clinical Informatics*. 13

LAUREN P., QU G., ZHANG F. & LENDASSE A. (2017). Discriminant document embeddings with an extreme learning machine for classifying clinical narratives. *Neurocomputing*, **277**, 129–138. Publisher: Elsevier. 35

LEAKE D. B. (1996). *Case-Based Reasoning: Experiences, lessons and future directions*. MIT press. 25

LEANING M. S., NG K. E. & CRAMP D. G. (1992). Decision support for patient management in oncology. *Medical Informatics = Medecine Et Informatique*, **17**(1), 35–46. 13

LEE J., MASLOVE D. M. & DUBIN J. A. (2015). Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PloS One*, **10**(5), e0127428. 27

LEE J., SUN J., WANG F., WANG S., JUN C.-H. & JIANG X. (2018a). Privacy-Preserving Patient Similarity Learning in a Federated Environment: Development and Analysis. *JMIR medical informatics*, **6**(2), e20. 26, 27

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. & KANG J. (2020). Biobert: A pretrained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240. 29

LEE W.-S., AHN S. M., CHUNG J.-W., KIM K. O., KWON K. A., KIM Y., SYM S., SHIN D., PARK I., LEE U. & BAEK J.-H. (2018b). Assessing Concordance With Watson for Oncology, a Cognitive Computing Decision Support System for Colon Cancer Treatment in Korea. *JCO clinical cancer informatics*, **2**, 1–8. 16

LERNER I., PARIS N. & TANNIER X. (2020). Terminologies augmented recurrent neural network model for clinical named entity recognition. *Journal of Biomedical Informatics*, **102**, 103356. 33

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880, Stroudsburg, PA, USA: Association for Computational Linguistics. 29

LI F., JIN Y., LIU W., RAWAT B., CAI P. & YU H. (2019). Fine-tuning bidirectional encoder representations from transformers (BERT)-based models on large-scale electronic health record notes: an empirical study. *Journal of Medical Internet Research*, **21**(11), e14830. 30

LI J., LIN Y., ZHAO P., LIU W., CAI L., SUN J., ZHAO L., YANG Z., SONG H., LV H. & WANG Z. (2022). Automatic text classification of actionable radiology reports of tinnitus patients using bidirectional encoder representations from transformer (bert) and in-domain pre-training (idpt). *BMC Med Inform Decis Mak*, **22**(1), 200. 64, 87, 125

LI L., CHENG W.-Y., GLICKSBERG B. S., GOTTESMAN O., TAMLER R., CHEN R., BOTTINGER E. P. & DUDLEY J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science Translational Medicine*, **7**(311), 311ra174. 27

LI W., QI K., CHEN W. & ZHOU Y. (2021). Unified batch all triplet loss for visible-infrared person re-identification. *CoRR*, **abs/2103.04607**. 98

LI Z., LIAN Y., MA X., ZHANG X. & LI C. (2020). Bio-semantic relation extraction with attention-based external knowledge reinforcement. *BMC Bioinformatics*, **21**(1), 213. 30

LIASHCHYNSKYI P. & LIASHCHYNSKYI P. (2019). Grid search, random search, genetic algorithm: A big comparison for nas. 66

LIN D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics-*, volume 2, p. 768–774, Morristown, NJ, USA: Association for Computational Linguistics. 30

LIU X., FAES L., CALVERT M. J., DENNISTON A. K. & CONSORT/SPIRIT-AI EXTENSION GROUP (2019a). Extension of the CONSORT and SPIRIT statements. *Lancet (London, England)*, **394**(10205), 1225. 14

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019b). Roberta: A robustly optimized bert pretraining approach. 29, 64

LOBACH D. F. & HAMMOND W. E. (1997). Computerized decision support based on a clinical practice guideline improves compliance with care standards. *The American Journal of Medicine*, **102**(1), 89–98. 13

LOH W.-Y. (2011). *Classification and regression trees*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 36

LU W. & ROTH D. (2015). Joint mention extraction and classification with mention hypergraphs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 857–867, Stroudsburg, PA, USA: Association for Computational Linguistics. 30

LUAN Y., WADDEN D., HE L., SHAH A., OSTENDORF M. & HAJISHIRZI H. (2019). A general framework for information extraction using dynamic span graphs. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, p. 3036–3046. 30

LUCAS P. (1993). The representation of medical reasoning models in resolution-based theorem provers. *Artificial intelligence in medicine*, **5**(5), 395–414. 24

LÖPER D., KLETTKE M., BRUDER I. & HEUER A. (2013). Enabling flexible integration of healthcare information using the entity-attribute-value storage model. *Health Information Science and Systems*, **1**, 9. 43

MACLAUGHLIN K. L., KESSLER M. E., KOMANDUR ELAYAVILLI R., HICKEY B. C., SCHEITEL M. R., WAGHOLIKAR K. B., LIU H., KREMERS W. K. & CHAUDHRY R. (2018). Impact of Patient Reminders on Papanicolaou Test Completion for High-Risk Patients Identified by a Clinical Decision Support System. *Journal of Women's Health (2002)*, **27**(5), 569–574. 16

MAGRATH M., YANG E., AHN C., MAYORGA C. A., GOPAL P., MURPHY C. C., GUPTA S., AGRAWAL D., HALM E. A., BORTON E. K., SKINNER C. S. & SINGAL A. G. (2018). Impact of a Clinical Decision Support System on Guideline Adherence of Surveillance Recommendations for Colonoscopy After Polypectomy. *Journal of the National Comprehensive Cancer Network: JNCCN*, **16**(11), 1321–1328. 13

MARAFINO B., DAVIES J., BARDACH N., DEAN M., DUDLEY R. & BOSCARDIN J. (2014). N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *Journal of the American Medical Informatics Association*, **21**(5), 871–875. 64

MARTIN L., MULLER B., ORTIZ SUAREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Stroudsburg, PA, USA: Association for Computational Linguistics. 29

MARTINEZ D., ANANDA-RAJAH M. R., SUOMINEN H., SLAVIN M. A., THURSKY K. A. & CAVEDON L. (2015). Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans. *Journal of Biomedical Informatics*, **53**, 251–260. 35

MASINO A. J., GRUNDMEIER R. W., PENNINGTON J. W., GERMILLER J. A. & CRENSHAW E. B. (2016). Temporal bone radiology report classification using open source machine learning and natural language processing libraries. *BMC medical informatics and decision making*, **16**(1), 65. Publisher: BioMed Central. 35

MAZO C., KEARNS C., MOONEY C. & GALLAGHER W. M. (2020). Clinical Decision Support Systems in Breast Cancer: A Systematic Review. *Cancers*, **12**(2), 369. 17, 76

MENGGE X., YU B., ZHANG Z., LIU T., ZHANG Y. & WANG B. (2020). Coarse-to-fine pre-training for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6345–6354: Association for Computational Linguistics. 30

MEYSTRE S. M., HEIDER P. M., KIM Y., ARUCH D. B. & BRITTEN C. D. (2019). Automatic trial eligibility surveillance based on unstructured clinical data. *International Journal of Medical Informatics*, **129**, 13–19. 26

MICHAELSON J. S., CHEN L. L., BUSH D., FONG A., SMITH B. & YOUNGER J. (2011). Improved web-based calculators for predicting breast carcinoma outcomes. *Breast Cancer Research and Treatment*, **128**(3), 827–835. 15

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *ArXiv*, p. 1–12. 29, 64

MING C., VIASSOLO V., PROBST-HENSCH N., DINOV I. D., CHAPPUIS P. O. & KATAPODI M. C. (2020). Machine learning-based lifetime breast cancer risk reclassification compared with the BOADICEA model: impact on screening recommendations. *British Journal of Cancer*, **123**(5), 860–867. 14

MIWA M., THOMPSON P., KORKONTZELOS I. & ANANIADOU S. (2014). Comparable study of event extraction in newswire and biomedical domains. In *COLING 2014 - 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers*, p. 2270–2279. 32

MIWA M., THOMPSON P., MCNAUGHT J., KELL D. B. & ANANIADOU S. (2012). Extracting semantically enriched events from biomedical literature. *BMC bioinformatics*, **13**, 108. 32

MOORE C., FARRAG A. & ASHKIN E. (2017). Using natural language processing to extract abnormal results from cancer screening reports. *Journal of patient safety*, **13**(3), 138. 32

MOUAZER A., SEDKI K., TSOPRA R. & LAMY J. (2021). Visual comparison of guidelines: Method and application to potentially inappropriate medication lists. In *Public Health and Informatics*, p. 248–252: IOS Press. 21

MUGGIA F. M. (1984). Multidisciplinary considerations in cancer treatment: Origin and scope. *International Journal of Radiation Oncology*Biology*Physics*, **10**, 31–33. 1

MUIS A. O. & LU W. (2017). Labeling gaps between words: Recognizing overlapping mentions with mention separators. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2608–2618, Stroudsburg, PA, USA: Association for Computational Linguistics. 30

MUJTABA G., SHUIB L., IDRIS N., HOO W. L., RAJ R. G., KHOWAJA K., SHAIKH K. & NWEKE H. F. (2019). Clinical text classification research trends: Systematic literature review and open issues. *Expert Systems with Applications*, **116**, 494–520. 34

MUJTABA G., SHUIB L., RAJ R. G., RAJANDRAM R., SHAIKH K. & AL-GARADI M. A. (2018). Classification of forensic autopsy reports through conceptual graph-based document representation model. *Journal of Biomedical Informatics*, **82**, 88–105. Publisher: Elsevier. 34, 36

NADEAU D. & SEKINE S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, **30**(1), 3–26. 29

NADKARNI P. M., MARENCO L., CHEN R., SKOUFOS E., SHEPHERD G. & MILLER P. (1999). Organization of heterogeneous scientific data using the EAV/CR representation. *Journal of the American Medical Informatics Association: JAMIA*, **6**(6), 478–493. 43

NAGENDRAN M., CHEN Y., LOVEJOY C. A., GORDON A. C., KOMOROWSKI M., HARVEY H., TOPOL E. J., IOANNIDIS J. P. A., COLLINS G. S. & MARUTHAPPU M. (2020). Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ (Clinical research ed.)*, **368**, m689. 14

NÊVÊOL A., DALIANIS H., VELUPILLAI S., SAVOVA G. & ZWEIGENBAUM P. (2018). Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of Biomedical Semantics*, **9**(1), 12. 32

NOLL R., SCHAAF J. & STORF H. (2022). The Use of Computer-Assisted Case-Based Reasoning to Support Clinical Decision-Making – A Scoping Review. In M. T. KEANE & N. WIRATUNGA, Coordinateurs, *Case-Based Reasoning Research and Development*, Lecture Notes in Computer Science, p. 395–409, Cham: Springer International Publishing. 99

NORGEOT B., QUER G., BEAULIEU-JONES B. K., TORKAMANI A., DIAS R., GIANFRANCESCO M., ARNAOUT R., KOHANE I. S., SARIA S., TOPOL E., OBERMEYER Z., YU B. & BUTTE A. J. (2020). Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature Medicine*, **26**(9), 1320–1324. 3, 14, 114

NOVIKAVA N., REDJDAL A., BOUAUD J. & SEROUSSI B. (2023). Clinical Decision Support Systems Applied to the Management of Breast Cancer Patients: A Scoping Review. *Studies in Health Technology and Informatics*, **305**, 353–356. 8, 15

NUNES T., CAMPOS D., MATOS S. & OLIVEIRA J. L. (2013). BeCAS: biomedical concept recognition services and visualization. *Bioinformatics*, **29**(15), 1915–1916. 31

OEI R. W., FANG H. S. A., TAN W.-Y., HSU W., LEE M.-L. & TAN N.-C. (2021). Using Domain Knowledge and Data-Driven Insights for Patient Similarity Analytics. *Journal of Personalized Medicine*, **11**(8), 699. 26

OHNO-MACHADO L., GENNARI J. H., MURPHY S. N., JAIN N. L., TU S. W., OLIVER D. E., PATTISON-GORDON E., GREENES R. A., SHORTLIFFE E. H. & BARNETT G. O. (1998). The guideline interchange format: a model for representing guidelines. *Journal of the American Medical Informatics Association: JAMIA*, **5**(4), 357–372. 18

ONG T. C., KAHN M. G., KWAN B. M. & AL E. (2017). Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading. *BMC Med Inform Decis Mak*, **17**, 134. 16

PAI S. & BADER G. D. (2018). Patient Similarity Networks for Precision Medicine. *Journal of Molecular Biology*, **430**(18), 2924–2938. 26, 27, 104, 129

PAI S., HUI S., ISSERLIN R., SHAH M. A., KAKA H. & BADER G. D. (2019). netDx: interpretable patient classification using integrated patient similarity networks. *Molecular Systems Biology*, **15**(3), e8497. 26, 27, 104, 129

PAL S. K. & SHIU S. C. (2004). *Foundations of Soft Case-Based Reasoning*. John Wiley & Sons. 25

PANAHIAZAR M., TASLIMITEHRANI V., PEREIRA N. L. & PATHAK J. (2015). Using EHRs for Heart Failure Therapy Recommendation Using Multidimensional Patient Similarity Analytics. *Studies in Health Technology and Informatics*, **210**, 369–373. 26, 27, 104, 129

PAPPAGARI R., ŻELASKO P., VILLALBA J., CARMIEL Y. & DEHAK N. (2019). Hierarchical transformers for long document classification. *arXiv*. 67

PARIMBELLI E., MARINI S., SACCHI L. & BELLAZZI R. (2018). Patient similarity for precision medicine: A systematic review. *Journal of Biomedical Informatics*, **83**, 87–96. 26

PARSHUTIN S. & KIRSHNERS A. (2013). Research on clinical decision support systems development for atrophic gastritis screening. *Expert Systems with Applications*, **40**(15), 6041–6046. 12

PATKAR V., ACOSTA D., DAVIDSON T., JONES A., FOX J. & KESHTGAR M. (2012). Using computerised decision support to improve compliance of cancer multidisciplinary meetings with evidence-based guidance. *BMJ open*, **2**(3), e000439. 16

PEDERSEN T., PAKHOMOV S., PATWARDHAN S. & CHUTE C. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *J Biomed Inform*, **40**(3), 288–299. 20

PELAYO S., BOUAUD J., BLANCAFORT C., LAMY J.-B., SEKAR B. D., LARBURU N., MURO N., RIBATE A. U., BELLOSO J., VALDERAS G., GUARDIOLA S., NGO C., TEIXEIRA L., GUÉZENNEC G. & SEROUSSI B. (2020). Preliminary Qualitative and Quantitative Evaluation of DESIREE, a Decision Support Platform for the Management of Primary Breast Cancer Patients. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, **2020**, 1012–1021. 16, 76

PELEG M. (2013). Computer-interpretable clinical guidelines: A methodological review. *Journal of Biomedical Informatics*, **46**(4), 744–763. 18

PELEG M., TU S., BURY J., CICCARESE P., FOX J., GREENES R. A., HALL R., JOHNSON P. D., JONES N., KUMAR A., MIKSCH S., QUAGLINI S., SEYFANG A., SHORTLIFFE E. H. & STEFANELLI M. (2003). Comparing Computer-interpretable Guideline Models: A Case-study Approach. *Journal of the American Medical Informatics Association*, **10**(1), 52–68. 18

PENG Y., YAN S. & LU Z. (2019). Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *arXiv preprint arXiv:1906.05474*. 30

PENNINGTON J., SOCHER R. & MANNING C. D. (2014a). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543. 29

PENNINGTON J., SOCHER R. & MANNING C. D. (2014b). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543. 68

PEREIRA S., NÉVÉOL A., KERDELHUÉ G. & ET AL. (2008). Using multi-terminology indexing for the assignment of mesh descriptors to health resources in a french online catalogue. *AMIA symp*, p. 586–590. 33, 65

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, p. 2227–2237. 29

PINEDA A. L., YE Y., VISWESWARAN S., COOPER G. F., WAGNER M. M. & TSUI F. R. (2015). Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. *Journal of Biomedical Informatics*, **58**, 60–69. 34, 36

PLANTIER M., HAVET N., DURAND T., CAQUOT N., AMAZ C., BIRON P., PHILIP I. & PERRIER L. (2017). Does adoption of electronic health records improve the quality of care management in France? Results from the French e-SI (PREPS-SIPS) study. *International Journal of Medical Informatics*, **102**, 156–165. 40

POPESCU M.-C., BALAS V., PERESCU-POPESCU L. & MASTORAKIS N. (2009). Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, **8**. 66

POTTER D., BROTHERS R., KOLACEVSKI A., KOSKIMAKI J. E., MCNUTT A., MILLER R. S., NAGDA J., NAIR A., RUBINSTEIN W. S., STEWART A. K., TRIEB I. J. & KOMATSOULIS G. A. (2020). Development of CancerLinQ, a Health Information Learning Platform From Multiple Electronic Health Record Systems to Support Improved Quality of Care. *JCO clinical cancer informatics*, **4**, 929–937. 16

POURNIK O., AHMAD B., DESPOTOU G., LIM CHOI KEUNG S. N., MOHAMAD Y., GAPPA H., LALECI ERTURKMEN G. B., YUKSEL M., GENCTURK M., SCHMIDT-BARZYNSKI W., STEINHOFF A., ROBBINS T., KYROU I., RANDEVA H., AYADI J., ARVANITIS T. N., ALCANTUD CÓRCOLES R., ABIZANDA P., LE K., GÓMEZ JIMÉNEZ E., AVENDAÑO CÉSPEDES A., KABA E. & MUIR H. (2023). CAREPATH methodology for development of computer interpretable, integrated clinical guidelines. In *Proceedings of the 10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*, DSAI '22, p. 7–11, New York, NY, USA: Association for Computing Machinery. 18

QIU J. X., YOON H. J., FEARN P. A. & TOURASSI G. D. (2018). Deep learning for automated extraction of primary sites from cancer pathology reports. *IEEE Journal of Biomedical and Health Informatics*, **22**(1), 244–251. 32

QIU X., SUN T., XU Y., SHAO Y., DAI N. & HUANG X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, **63**(10), 1872–1897. 29

QUAGLINI S. (2008). Compliance with clinical practice guidelines. *Studies in Health Technology and Informatics*, **139**, 160–179. 17

QUAGLINI S. & CICCARESE P. (2006). Models for guideline representation. *Neurological Sciences: Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, **27 Suppl 3**, S240–244. 18

RADFORD A., NARASIMHAN K., SALIMANS T. & SUTSKEVER I. (2018). Improving language understanding by generative pre-training. *OpenAI Blog*, p. 1–12. 29

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2020). *Language Models are Unsupervised Multitask Learners*. Rapport interne, OpenAI. 29

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, **21**, 1–67. 29

RAGHAVAN P., CHEN J. L., FOSLER-LUSSIER E. & LAI A. M. (2014). How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? In *AMIA Joint Summits on Translational Science proceedings*, p. 218–223. 3, 40, 115, 118

RAK R., ROWLEY A., BLACK W. & ANANIADOU S. (2012). Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database*, **2012**, bas010. 31

RAMCHANDRAN K. J., SHEGA J. W., VON ROENN J., SCHUMACHER M., SZMUILOWICZ E., RADEMAKER A., WEITNER B. B., LOFTUS P. D., CHU I. M. & WEITZMAN S. (2013). A predictive model to identify hospitalized cancer patients at risk for 30-day mortality based on admission criteria via the electronic medical record. *Cancer*, **119**(11), 2074–2080. 14

RANI G. J. J., GLADIS D. & MAMMEN J. (2015). Classification and prediction of breast cancer data derived using natural language processing. In *Proceedings of the third international symposium on women in computing and informatics (Wci-2015)*, p. 250–255. 34, 36

RATINOV L. & ROTH D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning - CoNLL '09*, p. 147, Boulder, Colorado: Association for Computational Linguistics. 157

RAU L. (1990). Extracting company names from text. In *Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application*, volume 1, p. 29–32: IEEE Comput. Soc. Press. 30

REDJDAL A., BOUAUD J., GLIGOROV J. & SEROUSSI B. (2021a). Are semantic annotators able to extract relevant complexity-related concepts from clinical notes? *Stud Health Technol Inform*, **287**, 153–157. 8, 65

REDJDAL A., BOUAUD J., GLIGOROV J. & SEROUSSI B. (2022a). Comparison of MetaMap, cTAKES, SIFR, and ECMT to Annotate Breast Cancer Patient Summaries. *Studies in Health Technology and Informatics*, **290**, 187–191. 8, 40, 65

REDJDAL A., BOUAUD J., GLIGOROV J. & SEROUSSI B. (2022b). Using machine learning and deep learning methods to predict the complexity of breast cancer cases. *Stud Health Technol Inform*, **294**, 78–82. 8, 66, 67, 126

REDJDAL A., BOUAUD J., GUÉZENNEC G., GLIGOROV J. & SEROUSSI B. (2021b). Reusing decisions made with one decision support system to assess a second decision support system: Introducing the notion of complex cases. 4, 116

REDJDAL A., BOUAUD J., GUÉZENNEC G., GLIGOROV J. & SEROUSSI B. (2021c). Reusing Decisions Made with One Decision Support System to Assess a Second Decision Support System: Introducing the Notion of Complex Cases. *Studies in Health Technology and Informatics*, **281**, 649–653. 8, 115

RICCI-CABELLO I., CARVALLO-CASTAÑEDA D., VÁSQUEZ-MEJÍA A., ALONSO-COELLO P., SAZ-PARKINSON Z., PARMELLI E., MORGANO G. P., RIGAU D., SOLÀ I., NEAMTIU L. & NIÑO-DE GUZMÁN E. (2023). Characteristics and impact of interventions to support healthcare providers' compliance with guideline recommendations for breast cancer: a systematic literature review. *Implementation science: IS*, **18**(1), 17. 76

RILOFF E. & JONES R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the National Conference on Artificial Intelligence*, number 1032, p. 474–479. 30

RINDFLESCH T. C. & FISZMAN M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, **36**(6), 462–477. 32

ROUSSEAU A.-L., BAUDELAIRE C. & RIERA K. (2020). Doctor GPT-3: hype or reality? 31

SAHOO S., KRINGS G., CHEN Y.-Y., CARTER J. M., CHEN B., GUO H., HIBSHOOSH H., REISENBICHLER E., FAN F., WEI S., KHAZAI L., BALASSANIAN R., KLEIN M. E., SHAD S., VENTERS S. J., BOROWSKY A. D., SYMMANS W. F. & OCAL I. T. (2022). Standardizing Pathologic Evaluation of Breast Carcinoma After Neoadjuvant Chemotherapy. *Archives of Pathology & Laboratory Medicine*, **147**(5), 591–603. 15

SAKJI S., GICQUEL Q., PEREIRA S., KERGOURLAY I., PROUX D., DAR-MONI S. & METZGER M. (2010). Evaluation of a french medical multi-terminology indexer for the manual annotation of natural language medical reports of healthcare-associated infections. *Stud Health Technol Inform*, **160**(Pt 1), 252–256. 33, 64, 65

SARKER A. & GONZALEZ G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, **53**, 196–207. 35

SAVOVA G. K., MASANZ J. J., OGREN P. V. & ET AL. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *J Am Med Inform Assoc*, **17**(5), 507–513. 32

SAVOVA G. K., TSEYTLIN E., FINAN S. & ET AL. (2017). Deepphe: A natural language processing system for extracting cancer phenotypes from clinical records. *Cancer Res*, **77**(21), e115–e118. 32

SCHIAPPA R., CONTU S., CULIE D., THAMPHYA B., CHATEAU Y., GAL J., BAILLEUX C., HAUDEBOURG J., FERRERO J.-M., BARRANGER E. & CHAMOREY E. (2022). RUBY: Natural Language Processing of French Electronic Medical Records for Breast Cancer Research. *JCO Clinical Cancer Informatics*, (6), e2100199. Publisher: Wolters Kluwer. 33, 60

SCHLEGEL D., GORDON K., GAUDIOSO C. & PELEG M. (2019). Clinical tractor: A framework for automatic natural language understanding of clinical practice guidelines. *AMIA Annu Symp Proc AMIA Symp*, **2019**, 784–793. 20

SCHMIDT R., MONTANI S., BELLAZZI R., PORTINALE L. & GIERL L. (2001). Case-based reasoning for medical knowledge-based systems. *International Journal of Medical Informatics*, **64**(2), 355–367. 25

SCHMIDT R., POLLWEIN B. & GIERL L. (1999). Experiences with case-based reasoning methods and prototypes for medical knowledge-based systems. *Artificial Intelligence in Medicine, Lecture Notes in Computer Science*, **1620**, 124–132. 25

SCHROFF F., KALENICHENKO D. & PHILBIN J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, p. 815–823. 28

SCOTT J. G., HARRISON L. B. & TORRES-ROCA J. F. (2017). Genomic biomarkers for precision radiation medicine - Authors' reply. *The Lancet. Oncology*, **18**(5), e239. 14

SEDGHI E., WEBER J. H., THOMO A., BIBOK M. & PENN A. M. (2016). A new approach to distinguish migraine from stroke by mining structured and unstructured clinical data-sources. *Network Modeling Analysis in Health Informatics and Bioinformatics*, **5**, 30. 35

SEROUSSI B., BLASZKA-JAULERRY B., ZELEK L., LEFRANC J.-P., CONFORTI R., SPANO J.-P., ROUSSEAU A. & BOUAUD J. (2012a). Accuracy of clinical data entry when using a computerized decision support system: a case study with OncoDoc2. *Studies in Health Technology and Informatics*, **180**, 472–476. 4, 17, 116

SEROUSSI B., BOUAUD J. & ANTOINE E. C. (2001). ONCODOC: a successful experiment of computer-supported guideline development and implementation in the treatment of breast cancer. *Artificial Intelligence in Medicine*, **22**(1), 43–64. 15

SEROUSSI B., BOUAUD J., GLIGOROV J. & UZAN S. (2007). Supporting multidisciplinary staff meetings for guideline-based breast cancer management: a study with OncoDoc2. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, **2007**, 656–660. 15, 16

SEROUSSI B., LAMY J.-B., MURO N., LARBURU N., SEKAR B. D., GUÉZENNEC G. & BOUAUD J. (2018). Implementing Guideline-Based, Experience-Based, and Case-Based Approaches to Enrich Decision Support for the Management of Breast Cancer Patients in the DESIREE Project. *Studies in Health Technology and Informatics*, **255**, 190–194. 16

SEROUSSI B., SOULET A., MESSAI N., LAOUÉNAN C., MENTRÉ F. & BOUAUD J. (2012b). Patient clinical profiles associated with physician non-compliance despite the use of a guideline-based decision support system: a case study with OncoDoc2 using data mining techniques. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, **2012**, 828–837. 13

SHAH N. H., JONQUET C., CHIANG A. P., BUTTE A. J., CHEN R. & MUSEN M. A. (2009). Ontology-driven indexing of public datasets for translational bioinformatics. *BMC bioinformatics*, **10 Suppl 2**(Suppl 2), S1. 31

SHE Y., JIN Z., WU J., DENG J., ZHANG L., SU H., JIANG G., LIU H., XIE D., CAO N., REN Y. & CHEN C. (2020). Development and Validation of a Deep Learning Model for Non-Small Cell Lung Cancer Survival. *JAMA network open*, **3**(6), e205842. 14

SHIN B., CHOKSHI F. H., LEE T. & CHOI J. D. (2017). Classification of radiology reports using neural attention models. In *Proceeding of the 2017 international joint conference on neural networks (IJCNN)*, p. 4363–4370: IEEE. 34, 36

SHORTLIFFE E. H. (1986). Update on ONCOCIN: a chemotherapy advisor for clinical oncology. *Medical Informatics = Medecine Et Informatique*, **11**(1), 19–21. 13

SIMOUDIS E. & MILLER J. (1990). Validated retrieval in case-based reasoning. In *Proceedings of AAAI*, p. 310–315. 25

SOMASHEKHAR S. P., SEPÚLVEDA M.-J., PUGLIELLI S., NORDEN A. D., SHORTLIFFE E. H., ROHIT KUMAR C., RAUTHAN A., ARUN KUMAR N., PATIL P., RHEE K. & RAMYA Y. (2018). Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, **29**(2), 418–423. 15, 16

SOUKUP T., GANDAMIHARDJA T. A. K., MCINERNEY S., GREEN J. S. A. & SEVDALIS N. (2019). Do multidisciplinary cancer care teams suffer decision-making fatigue: an observational, longitudinal team improvement study. *BMJ open*, **9**(5), e027303. 3, 63, 124

SOUKUP T., LAMB B. W., MORBI A., SHAH N. J., BALI A., ASHER V., GANDAMIHARDJA T., GIORDANO P., DARZI A., SEVDALIS N. & GREEN J. S. A. (2022). Cancer multidisciplinary team meetings: impact of logistical challenges on communication and decision-making. *BJS open*, **6**(4), zrac093. 2, 113

SOUKUP T., MORBI A., LAMB B. W., GANDAMIHARDJA T. A. K., HOGBEN K., NOYES K., SKOLARUS T. A., DARZI A., SEVDALIS N. & GREEN J. S. A. (2020). A measure of case complexity for streamlining workflow in multidisciplinary tumor boards: Mixed methods development and early validation of the MeDiC tool. *Cancer Medicine*, **9**(14), 5143–5154. 3, 63

SOYSAL E., WANG J., JIANG M., WU Y., PAKHOMOV S., LIU H. & XU H. (2018). CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association: JAMIA*, **25**(3), 331–336. 32

SPACKMAN K. A., CAMPBELL K. E. & CÔTÉ R. A. (1997). Snomed rt: a reference terminology for health care. *Proceedings of the AMIA Annual Fall Symposium*, **4**(SUPPL.), 640. 35

STEARNS M. Q., PRICE C., SPACKMAN K. A. & WANG A. Y. (2001). SNOMED clinical terms: overview of the development process and project status. In *Proceedings of the AMIA symposium*, p. 662: American Medical Informatics Association. 35

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012a). Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 102–107, Stroudsburg, PA, USA: Association for Computational Linguistics. 31

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012b). brat: a Web-based Tool for NLP-Assisted Text Annotation. p. 102–107. 44, 120

SUJANSKY W. (2001). Heterogeneous database integration in biomedicine. *J Biomed Inform*, **34**, 285–298. 16

SUN C., QIU X., XU Y. & HUANG X. (2019). How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, p. 194–206: Springer. 67

SUN R., LIMKIN E. J., VAKALOPOULOU M., DERCLE L., CHAMPIAT S., HAN S. R., VERLINGUE L., BRANDAO D., LANCIA A., AMMARI S., HOLLEBECQUE A., SCOAZEC J.-Y., MARABELLE A., MASSARD C., SORIA J.-C., ROBERT C., PARAGIOS N., DEUTSCH E. & FERTÉ C. (2018). A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *The Lancet. Oncology*, **19**(9), 1180–1191. 14

SUNDQVIST M., CHIQUET J. & RIGAILL G. (2020). Adjusting the adjusted Rand Index – A multinomial story. Issue: arXiv:2011.08708 arXiv: 2011.08708 [stat]. 99

SUNG H., FERLAY J., SIEGEL R. L., LAVERSANNE M., SOERJOMATARAM I., JEMAL A. & BRAY F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians*, **71**(3), 209–249. 1, 113

SUO Q., MA F., YUAN Y., HUAI M., ZHONG W., GAO J. & ZHANG A. (2018). Deep Patient Similarity Learning for Personalized Healthcare. *IEEE transactions on nanobioscience*, **17**(3), 219–227. 27

SUTTON R. T., PINCOCK D., BAUMGART D. C. & OTHERS (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med*, **3**, 17. 14

SYED Y. (2020). Oncotype dx breast recurrence score®: A review of its use in early-stage breast cancer. *Mol Diagn Ther*, **24**(5), 621–632. 70

SÉROUSSI B., SOULET A., SPANO J.-P., LEFRANC J.-P., COJEAN-ZELEK I., BLASZKA-JAULERRY B., ZELEK L., DURIEUX A., TOURNIGAND C., MESSAI N., ROUSSEAU A. & BOUAUD J. (2013). Which patients may benefit from the use of a decision support system to improve compliance of physician decisions with clinical practice guidelines: a case study with breast cancer involving data mining. *Studies in Health Technology and Informatics*, **192**, 534–538. 16

TAIRA R. K., SODERLAND S. G. & JAKOBOVITS R. M. (2001). Automatic structuring of radiology free-text reports. *Radiographics*, **21**(1), 237–245. 32

TAN Z., SHEN Y., ZHANG S., LU W. & ZHUANG Y. (2021). A sequence-to-set network for nested named entity recognition. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, p. 3936–3942: International Joint Conferences on Artificial Intelligence Organization. 30

TCHECHMEDJIEV A., ABDAOUI A., EMONET V. & ET AL. (2018). Sifr annotator: ontology-based semantic annotation of french biomedical text and clinical notes. *BMC Bioinformatics*, **19**(1), 405. 33

THOMPSON R. F., VALDES G., FULLER C. D., CARPENTER C. M., MORIN O., ANEJA S., LINDSAY W. D., AERTS H. J. W. L., AGRIMSON B., DEVILLE C., ROSENTHAL S. A., YU J. B. & THOMAS C. R. (2018). Artificial intelligence in radiation oncology: A specialty-wide disruptive transformation? *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology*, **129**(3), 421–426. 3, 14, 114

TIEDEMANN J. & THOTTINGAL S. (2020). OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, p. 479–480, Lisboa, Portugal. 65

TORRENTE M., SOUSA P. A., HERNÁNDEZ R., BLANCO M., CALVO V., COLLAZO A., GUERREIRO G. R., NÚÑEZ B., PIMENTAO J., SÁNCHEZ J. C., CAMPOS M., COSTABELLO L., NOVACEK V., MENASALVAS E., VIDAL M. E. & PROVENCIO M. (2022). An Artificial Intelligence-Based Tool for Data Analysis and Prognosis in Cancer Patients: Results from the Clarify Study. *Cancers*, **14**(16), 4041. 16

TRAN K. A., KONDRASHOVA O., BRADLEY A., WILLIAMS E. D., PEARSON J. V. & WADDELL N. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine*, **13**, 152. 14

TRAN T. T. T., NGHIEM S. V., LE V. T., QUAN T. T., NGUYEN V., YIP H. Y. & BODENREIDER O. (2020). Siamese KG-LSTM: A deep learning model for enriching UMLS Metathesaurus synonymy. *The ... International Conference on Knowledge and Systems Engineering. International Conference on Knowledge and Systems Engineering*, **2020**, 281–286. 98

TURIAN J., RATINOV L. & BENGIO Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, p. 384–394. 29

TUTUBALINA E., MIFTAHUTDINOV Z., NIKOLENKO S. & MALYKH V. (2018). Medical concept normalization in social media posts with recurrent neural networks. *J Biomed Inform*, **84**, 93–102. 20

UYSAL A. K. & GUNAL S. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems*, **36**, 226–235. 36

VAN AKEN B., PAPAIOANNOU J.-M., NAIK M. G., ELEFTHERIADIS G., NEJDL W., GERS F. A. & LÖSER A. (2022). This patient looks like that patient: Prototypical networks for interpretable diagnosis prediction from clinical text. 27

VAN HOEVE J., DE MUNCK L., OTTER R., DE VRIES J. & SIESLING S. (2014). Quality improvement by implementing an integrated oncological care pathway for breast cancer patients. *Breast (Edinburgh, Scotland)*, **23**(4), 364–370. 2, 113

VENOT A. (2013). *Informatique Médicale, e-Santé – Fondements et applications*. Springer: Springer. 17

VERBERNE C. J., NIJBOER C. H., DE BOCK G. H., GROSSMANN I., WIGGERS T. & HAVENGA K. (2012). Evaluation of the use of decision-support software in carcino-embryonic antigen (CEA)-based follow-up of patients with colorectal cancer. *BMC medical informatics and decision making*, **12**, 14. 13

VERBORGH R. & DE ROO J. (2015). Drawing Conclusions from Linked Data on the Web: The EYE Reasoner. *IEEE Software*, **32**(3), 23–27. Conference Name: IEEE Software. 23

VOIGT W. & TRAUTWEIN M. (2023). Improved guideline adherence in oncology through clinical decision-support systems: still hindered by current health IT infrastructures? *Current Opinion in Oncology*, **35**(1), 68–77. 2, 13, 17, 76, 114

WAJSBÜRT P. (2021). *Extraction and normalization of simple and structured entities in medical documents*. Theses, Sorbonne Université. 157, 158, 162

WAJSBÜRT P. (2021). *Extraction and normalization of simple and structured entities in medical documents*. phdthesis, Sorbonne Université. 33

WAJSBÜRT P., SARFATI A. & TANNIER X. (2021). Medical concept normalization in French using multilingual terminologies and contextual embeddings. *Journal of Biomedical Informatics*, **114**, 103684. 33

WAKS Z., GOLDBRAICH E., FARKASH A. & AL E. (2013). Analyzing the 'CareGap': assessing gaps in adherence to clinical guidelines in adult soft tissue sarcoma. *Stud Health Technol Inform*, **186**, 46–50. 13

WANG J., SHOU L., CHEN K. & CHEN G. (2020). Pyramid: A layered model for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5918–5928: Association for Computational Linguistics. 30

WANG L., CHEN X., ZHANG L., LI L., HUANG Y., SUN Y. & YUAN X. (2023). Artificial intelligence in clinical decision support systems for oncology. *International Journal of Medical Sciences*, **20**(1), 79–86. 12, 161

WANG L., CHU F. & XIE W. (2007). Accurate cancer classification using expressions of very few genes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **4**(1), 40–53. 64

WANG N., HUANG Y., LIU H., FEI X., WEI L., ZHAO X. & CHEN H. (2019). Measurement and application of patient similarity in personalized predictive modeling based on electronic medical records. *BioMedical Engineering OnLine*, **18**(1), 98. 26

WANG Y., COIERA E., RUNCIMAN W. & MAGRABI F. (2017). Using multiclass classification to automate the identification of patient safety incident reports by type and severity. *Bmc Medical Informatics and Decision Making*, **17**, 84. 35

WHITE R. W. & HORVITZ E. (2017). Evaluation of the Feasibility of Screening Patients for Early Signs of Lung Carcinoma in Web Search Logs. *JAMA oncology*, **3**(3), 398–401. 14

WIENEKE A. E., BOWLES E. J. A., CRONKITE D., WERNLI K. J., GAO H., CARRELL D. & BUIST D. S. M. (2015). Validation of natural language processing to extract breast cancer pathology procedures and results. *Journal of Pathology Informatics*, **6**, 38. 40

WISHART G. C., AZZATO E. M., GREENBERG D. C., RASHBASS J., KEARINS O., LAWRENCE G., CALDAS C. & PHAROAH P. D. P. (2010). PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast cancer research: BCR*, **12**(1), R1. 15

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R. & FUNTOWICZ M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*. 31

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M. *et al.* (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, p. 38–45: Association for Computational Linguistics. 29

XIE C., DUFFY A. G., BRAR G., FIORAVANTI S., MABRY-HRONES D., WALKER M., BONILLA C. M., WOOD B. J., CITRIN D. E., GIL RAMIREZ E. M., ESCORCIA F. E., REDD B., HERNANDEZ J. M., DAVIS J. L., GASMI B., KLEINER D., STEINBERG S. M., JONES J. C. & GRETEN T. F. (2020). Immune Checkpoint Blockade in Combination with Stereotactic Body Radiotherapy in Patients with Metastatic Pancreatic Ductal Adenocarcinoma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, **26**(10), 2318–2326. 14

XIONG W., DU J., WANG W. & STOYANOV V. (2020). Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *ICLR 2020: Eighth International Conference on Learning Representations*. 72

XU M., JIANG H. & WATCHARAWITTAYAKUL S. (2017). A local detection approach for named entity recognition and mention detection. In *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 1, p. 1237–1247. 30

XU Y., HOSNY A., ZELEZNIK R., PARMAR C., COROLLER T., FRANCO I., MAK R. H. & AERTS H. J. W. L. (2019). Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, **25**(11), 3266–3275. 14

YAN H., GUI T., DAI J., GUO Q., ZHANG Z. & QIU X. (2021). A unified generative framework for various ner subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 5808–5822, Stroudsburg, PA, USA: Association for Computational Linguistics. 30

YANG Y. & PEDERSEN J. (1997). A comparative study on feature selection in text categorization. In *ICML*, p. 412–420. 36

YANG Z., DAI Z., YANG Y., CARBONELL J., SALAKHUTDINOV R. & LE Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*. 29

YE Y., TSUI F., WAGNER M., ESPINO J. & LI Q. (2014). Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. *Journal of the American Medical Informatics Association*, **21**, 815–823. 34

ZHENG C., CAI Y., XU J., LEUNG H.-F. & XU G. (2019). A boundary-aware neural model for nested named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 357–366: Association for Computational Linguistics. 30

ZHU Y., YAN E. & WANG F. (2017). Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC Med Inform Decis Mak*, **17**(1), 95. 20

ZUCCON G., KHANNA S., NGUYEN A., BOYLE J., HAMLET M. & CAMERON M. (2015). Automatic detection of tweets reporting cases of influenza like illnesses in Australia. *Health Information Science and Systems*, **3**, S4. 35

ZUCCON G., WAGHOLIKAR A., NGUYEN A., BUTT L., CHU K., MARTIN S. & GREENSLADE J. (2013). Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the SNOMED CT ontology. In *AMIA Joint Summits on Translational Science Proceedings, 2013*, p. 300–304. 34, 36

ZWEIGENBAUM P., GROUIN C. & LAVERGNE T. (2016). Une catégorisation de fins de lignes non-supervisée (End-of-line classification with no supervision). In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 2 : TALN (Posters)*, p. 364–371, Paris, France: AFCP - ATALA. 49

Appendix

To maintain a smooth reading experience, detailed and supplementary materials are presented separately in the Appendices section. This ensures that the memoir's flow remains uninterrupted while providing interested readers easy access to valuable supporting information and data.

A Hybrid NER method

In chapter 3, we used rule-based techniques to perform named entity recognition. However, as we mentioned in the perspectives, we started hybridizing this method using a deep learning algorithm, the method is presented below, and an evaluation step is required to validate it.

A.1 Deep learning method for named entity recognition

Once we had the rule-based approach to do the annotation. We evaluated it on a corpus of breast cancer patient summaries, manually annotated by experts who took the pre-annotated files by the rule-based system, and corrected the errors done by the algorithm. After the evaluation, we determined the entities where the performance of the rule-based approach was poor and we trained a deep learning NER model to improve the performance of the algorithms. The data used to evaluate the rule-based system was the training corpus for the deep learning approach.

To implement this method we used EDS-NLP. Indeed, in addition to its rule-based pipeline components, EDS-NLP offers new trainable pipelines to fit and run machine learning models for classic biomedical information extraction tasks. The new eds.ner component allows to extract almost any named entity.

A.1.1 Model architecture

This method, developed by Wajsbürt (2021), utilizes a deep learning approach for named entity recognition. The model employs a token classification technique using the BIOUL tagging scheme, which consists of tags such as B (Begin), I (Inside), O (Outside), U (Unary), and L (Last). Each label in the model has its own tag sequence, allowing for the extraction of overlapping entities. (Ratinov & Roth, 2009) studied the BIOUL tag scheme and found it performs very well. This scheme encodes the end of entities and single words entities with specific tags E and S 1.

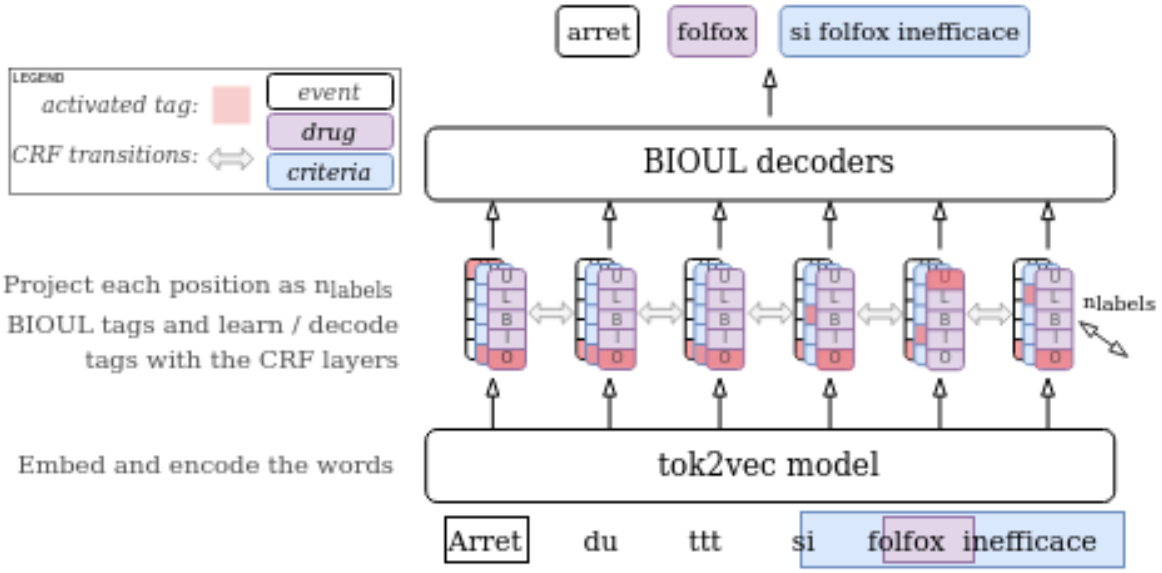


Figure A.1: Nested NER architecture by (Wajsbürt, 2021)

To enforce the tagging scheme and ensure valid tag sequences (e.g., preventing I from following O but only allowing B), the model utilizes a stack of CRF (Conditional Random Fields) layers. The CRF layers are employed during both training and prediction stages to capture the dependencies between the tags and ensure the output follows a valid sequence. The deep learning architecture used in the eds.ner component is designed to handle various types of named entities. It can extract flat entities similar to spaCy's EntityRecognizer, overlapping entities including those with different labels, and entities with ill-defined boundaries. This allows for more flexible and comprehensive named entity recognition.

In summary, the deep learning method employed by the eds.ner component utilizes a stack of CRF layers to enforce the BIOUL tagging scheme, enabling the extraction of various types of named entities. While it offers improved capabilities compared to spaCy's default NER pipeline, it still has limitations regarding nested entities of the same label.

B Update of the GL-DSS's knowledge base

In this section, we dive deeper into the results obtained in chapter 5, first, we provide an overview of the ontology, then we present all the rules identified to update the GL-DSS's knowledge base

B.1 Information on the BCKM ontology and the rule bases

Currently, the GL-DSS's conceptual model consists of 22 entities and a total of 394 attributes, distributed based on their value type as follows: 49% Booleans, 9% integers, 4% floats, 5% strings, 4% dates, and 33% hierarchical values. The BCKM ontology encompasses 1445 classes, 2305 axioms, 25 object properties, and 15 data properties. Notably, 658 classes are derived from the NCI thesaurus.

The breast cancer CPGs from AP-HP (France) that the GL-DSS was based on were published in 2016 as a 36-page document, providing diagnostic and therapeutic recommendations, distinguishing between surgery, chemotherapy, and endocrine therapy procedures. The rule base consists of 305 with a subset of generic 12 rules. Figure B.1 presents an attribute (bilateral breast cancer), that is an attribute of the patient entity and has a boolean value.

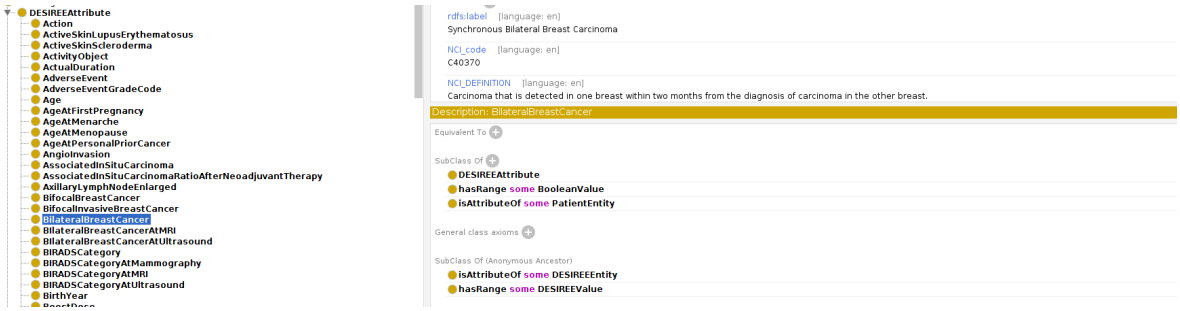


Figure B.1: Overview of the BCKM

B.2 Updates in the BCKM ontology and the rule bases

As presented in chapter 5, we identified updates in the GL-DSS's knowledge base, this led us to the modification, creation, and suppression of existing rules in the knowledge base, figure B.2 shows an example of comparison between GL-DSS and SENORIF recommendations with the MTB's decision.

A table that describes all the mappings is available in this link `https://drive.google.com/file/d/1HvkVD2FVZY4896tPyK92zi0zT0VsYOjH/view?usp=sharing` as the rules are written in French and also the guidelines, the table is written in French.

C Mapping of structured data with the ontology

As explained in section 5.2.2.1, We did a mapping of the structured data from chapter 3 to the BCKM ontology, the table that describes the mapping can be found using this link: `https://drive.google.com/file/d/1-WRRmU43T2iii2ZRsVsshkhn9PrLrGxZ/view?usp=sharing`.

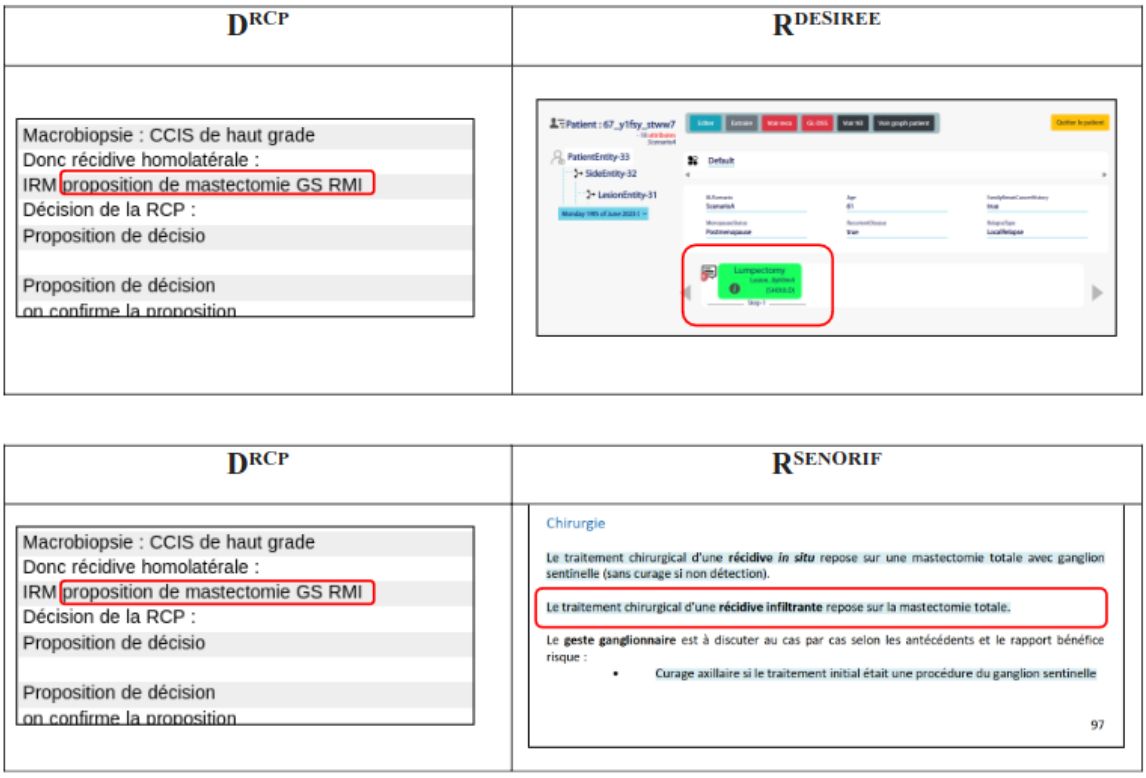


Figure B.2: Example of differences between guideline recommendations and MTB decision

List of Figures

List of tables