electronics

Article

Integrating Unstructured EHR Data Using an FHIR-Based System: A Case Study with Problem List Data and an FHIR IPS Model

Fouzia Amar *, Alain April and Alain Abran

Software Engineering and IT Department, École de Technologie Supérieure—ÉTS, Montréal, QC H3C 1K3, Canada; alain.april@etsmtl.ca (A.A.); alain.abran@etsmtl.ca (A.A.)

* Correspondence: fouzia.amar.1@ens.etsmtl.ca

Abstract

The patient problem list is a key component of an electronic health record (EHR) and must be accurate and accessible for all professionals involved in patient care. Unfortunately, such a list is mostly found in an unstructured text format, is not easily sharable across digital health systems, and lacks semantic interoperability. Natural language processing (NLP) techniques are widely used for clinical concept extraction, particularly for English text. However, in the Canadian context, the clinical notes in a patient problem list can also be found in French. This research presents a framework based on Fast Healthcare Interoperability Resources (FHIR) consisting of an NLP clinical pipeline and a rule-based approach to converting the textual patient problem list, including notes regarding allergies, into an FHIR model. The proposed approach considers concept modifiers to map to the International Patient Summary (IPS) FHIR model element. The main contributions of this research include the early detection of FHIR resources from unstructured data written in the French language and the design of a rule-based algorithm to identify and map extracted data to the appropriate FHIR resource attributes using an annotator. A primary evaluation of the resource tag which uses the rule-based method demonstrates the feasibility of the proposed model to facilitate semantic interoperability. The assessment was conducted using the French FRASIMED corpora.

Keywords: EHR; HL7; FHIR; interoperability; semantic; terminology; NLP; IPS; ML; SNOMED CT; rule-based

Academic Editors: Eleni Boumpa and Polat Goktas

Received: 7 September 2025
Revised: 7 October 2025
Accepted: 16 October 2025
Published: 22 October 2025

Citation: Amar, F.; April, A.; Abran, A. Integrating Unstructured EHR Data Using an FHIR-Based System: A Case Study with Problem List Data and an FHIR IPS Model. *Electronics* **2025**, *14*, 4134. https://doi.org/10.3390/electronics14214134

1. Introduction

The patient problem list is a key component of a patient's health record, identified as required in the International Patient Summary (IPS) [1]. It must be accurate and accessible to all involved in patient care. However, a patient problem list is often available mainly in an unstructured format, for example, clinical notes and pathology reports, etc. Such a list is typically difficult to share because it is unstructured.

Electronic health record (EHR) systems contain patient clinical information in various formats, including unstructured formats, which impede semantic interoperability. Facilitating the semantic interoperability of healthcare data, not only for primary care but also for secondary use (e.g., clinical analysis and medical research), using natural language processing (NLP) techniques and machine learning (ML) algorithms has gained popularity for dealing with unstructured data [2]. NLP techniques and tools are mostly available in the English language. However, in the Canadian context, and specifically in the province of Quebec, clinical notes can be written in either English or French. Since NLP techniques are language-specific and there is a need to convert clinical problems written in French into a standard-based model, the focus of this study is scenarios in French.

The IPS profile has been adopted by the International Organization for Standardization (ISO), Health Level Seven (HL7), and the Integrating the Healthcare Enterprise (IHE) initiative to facilitate semantic interoperability based on clinical data normalization [1]. The IPS dataset is minimal and non-exhaustive; it is a brief summary, but still clinically relevant, and is composed of FHIR resources classified into three categories: required, recommended, and optional [3]. The FHIR IPS specification describes the FHIR implementation of the IPS. It is an implementation guide defining how to create an IPS document using the FHIR standard. From an FHIR implementation perspective, the "Condition resource" is used to store information related to the patient problem list, the diagnosis, or other clinical concepts [4]. The patient problem list also includes information about allergies that need to be clearly identified and stored in the "Allergy Intolerance resource".

The research motivation of this study is to improve the data sharing of an existing free-text patient problem list written in French using standards that ensure semantic interoperability.

This research experiments with an FHIR-based framework consisting of an NLP clinical pipeline as well as a rule-based approach to converting a patient problem list written in French, including allergies, into an FHIR model. The proposed framework aims to transform the French unstructured text of the clinical problem list into an FHIR-based model using the IPS specification. The proposed approach considers concept modifiers while mapping to the IPS FHIR model elements. The feasibility of the proposed approach is evaluated using the FRASIMED dataset, an annotated corpus for clinical notes in French [5].

The remainder of this paper is structured as follows: Section 2 describes related works; Section 3 presents the research methodology, including research challenges, and the proposed approach; Section 4 considers the results that address the research challenges; and finally, Section 5 presents the discussion, followed by the conclusion regarding future work.

2. Related Work

A number of medication information extraction studies have proposed frameworks for transforming unstructured clinical text into FHIR models using an artificial intelligence natural language processing (NLP) pipeline composed of two main steps: (1) extracting concepts using different NLP techniques and (2) mapping the resulting concepts to a health domain ontology (e.g., UMLS, SNOMED CT) to ensure a shared understanding and enable data exchange where the meaning is preserved.

Durango et al. [6] presented different methods for automatically extracting valuable information using NLP techniques, most of which were applied to English texts. The most common open-source NLP tools highlighted in another survey [7] were (1) MetaMap, (2) MetaMap Lite, (3) Clinical Text Analysis and Knowledge Extraction System (cTAKES), and (4) Open Biomedical Annotator (OBA). Lee et al. [8] proposed the Biodirectional Encoder Representations from Transformers for Biomedical Text Mining (BIoBERT), a pre-trained language representation model, to extract valuable information from unstructured biomedical text.

NLP tools developed for the Unstructured Information Management Architecture (UIMA) architecture are widely used [9]. Their extraction process is based on several clinical NLP tools, such as cTAKES and MedXN. In another study specific to medication data [10], NLP-based mapping rules were used to convert unstructured data into an FHIR model, and the proposed model was applied to the FHIR resource "MedicationStatement".

A more general framework proposed by Hong et al. [11], referred to as NLP2FHIR, is an FHIR-based clinical data normalization pipeline mainly comprising an NLP pipeline with mapping and normalization rules that includes a module for integrating structured data. The NLP2FHIR clinical data normalization pipeline was used and extended by Liu et al. [12]. It includes a use-case scenario from an obesity database that applies deep learning models to text classification of NLP2FHIR outputs for analytics.

To customize NLP pipelines, Soysal et al. [13] present an NLP toolkit named Clinical Language Annotation, Modeling, and Processing (CLAMP) that enables non-NLP expert users to customize their NLP pipelines via a graphical user interface. Wang et al. [14] used the CLAMP toolkit with UMLS to automatically extract opioid information from free text and map it to an FHIR.

Peterson et al. [15] converted free text into FHIR using NLP models, namely Bidirectional Encoder Representations from Transformers (BERT) and UMLS, to map extracted concepts into FHIR models. A neural network classification model allows the concept to be mapped to the corresponding element in the "Condition resource", which includes the patient problem list.

In another study, Peterson and Liu [16] converted unstructured problem descriptions into SNOMED CT expressions using a deep learning method for relation identification between concepts and problem phrases. The concept extraction process from UMLS was performed using the MetaMap tool, a named-entity recognition tool developed by the National Library of Medicine (NLM).

The published results have shown that patient problem lists require extraction along with their context. For instance, Wu et al. [17] proposed an open-source semantic search tool to extract concepts from UMLS, including contextualized mentions (e.g., negation, temporality, and experiencer). The retrieval process is based on the NLP pipeline, which focuses on annotating the UMLS concepts in clinical notes.

SNOMED CT and related SNOMED tools [18] are widely used in English to facilitate semantic interoperability, mainly with rule-based approaches. Studies dealing with other languages, including French, remain limited. Most studies mapping to terminology use UMLS; however, there is no easy solution for aligning unstructured text with SNOMED CT. The study by Gaudet-Blavignac et al. [7] highlighted the target language, which was mainly English, while the other languages mentioned were Swedish, Czech, and Chinese. In the literature, the reported mapping methods are mainly rule-based (70%), manual (14%), hybrid (11%), and machine learning (5%) [7].

Almost all NLP techniques for the clinical field are applied to text written in English. Since NLP techniques are dependent on language, there is still research to be undertaken for other languages, such as French. A recent survey of the NLP techniques available in languages other than English for extraction and name entity recognition purposes confirmed the absence of a French pre-trained medical model [19]. Although French is an important language in UMLS, only 4% of all available concepts are linked to at least one label term in French. In the literature, one strategy for addressing this issue is to first perform a translation into English to then find a corresponding concept [19].

From this related work, we can see that no framework has addressed the semantic interoperability of a patient problem list using the FHIR format for unstructured text in French. The survey in [19] also confirmed that there is a lack of annotated datasets and models for languages other than English, including French.

3. Research Methodology

3.1. Problem Statement

The research motivation of this study is to improve the data sharing of an existing free-text patient problem list written in French using standards that ensure a common understanding (i.e., semantics) between the systems involved (i.e., allowing for better interoperability). This research focuses on converting the patient problem list and detecting allergies/intolerances from French unstructured text into an FHIR model.

This study proposes an FHIR-based framework in the Canadian context to address the challenges listed in Table 1.

Table 1. Research challenges.

ID	Category	Challenge Description
1.	Data format	
1.1		Information related to the patient problem list is mainly in unstructured format.
1.2		Most reports are in PDF file format.
2.	Language	
2.1		Clinical notes in Canada, and other French countries, are either in English or French.
2.2		NLP models and techniques are language-dependent. Selecting the appropriate NLP pipeline requires prior identification of the language used.
2.3		Most NLP tools are for English text. There is a major need in other languages, including French, which is largely used in Quebec, for the interoperability of the patient problem list.
3.	Context and modifiers	
3.1		The patient problem list may be related to an allergy/intolerance, a diagnosis, or other types of related clinical conditions. It is important to distinguish between these items to ensure correct mapping to the FHIR elements.
3.2		The proposed framework needs to consider that the extracted condition may be in a negation context.
3.3		The extracted condition may be related to the patient or their family members.
3.4		The extracted condition may be confirmed or only a hypothesis.
3.5		The extracted condition may be active or resolved (historical).
4.	Standard/guidelines	
4.1		A standard (e.g., SNOMED CT) must be used to ensure semantic interoperability or common understanding and interpretability.
5.	Condition type	
		The patient problem list may be related to an allergy or another type of health conditions. Allergies need to be distinguished from other condition types.

3.2. Method

The system architecture of the proposed framework consists of two stages (Figure 1):

1. Identification of the section and resource tags—executed only once (Stage 1);
2. Data processing—executed for every clinical note file (Stage 2).

- Stage 1: Tag Identification

Clinical notes are typically split into sections such as “diagnosis” or “physical exam”. Knowing the section from which the concept is extracted is an important semantic indicator. Therefore, section recognition is a fundamental step in NLP pipelines for unstructured data [20,21]. Furthermore, since allergies also belong to the list of clinical problems (the same section) and may be stored in different FHIR resources (Condition vs. Allergy

Intolerance), we seek a way to recognize this by using an allergy resource tag. Stage 1 involves four steps:

1. The identification of sections related to the patient problem list and diagnosis: The list of section tags in [21] with 6773 items was used as the starting point to identify and extract section candidates that may contain information related to the patient problem list and diagnosis. The resulting list was then optimized by eliminating duplicates and semantic matching using a pivot term. For example, principal diagnosis and secondary diagnosis were replaced by diagnosis (Figure 2).
2. Translation: Next, the optimized list was translated into French using two tools: DeepL (free version) [22] and the French medical dictionary (2025 free version) [23].
3. Data augmentation: The list was enriched further with ChatGPT (free version GPT-3.5) synonyms [24] using insights from several studies that have used synonyms for data augmentation [25–27]. All relevant synonyms were added to our list of section candidates.
4. Classification to categories: The final French tag list includes three main categories: tags related to allergies (50 elements), diagnosis (106 elements), and other conditions (156 elements) (see Supplementary Materials)

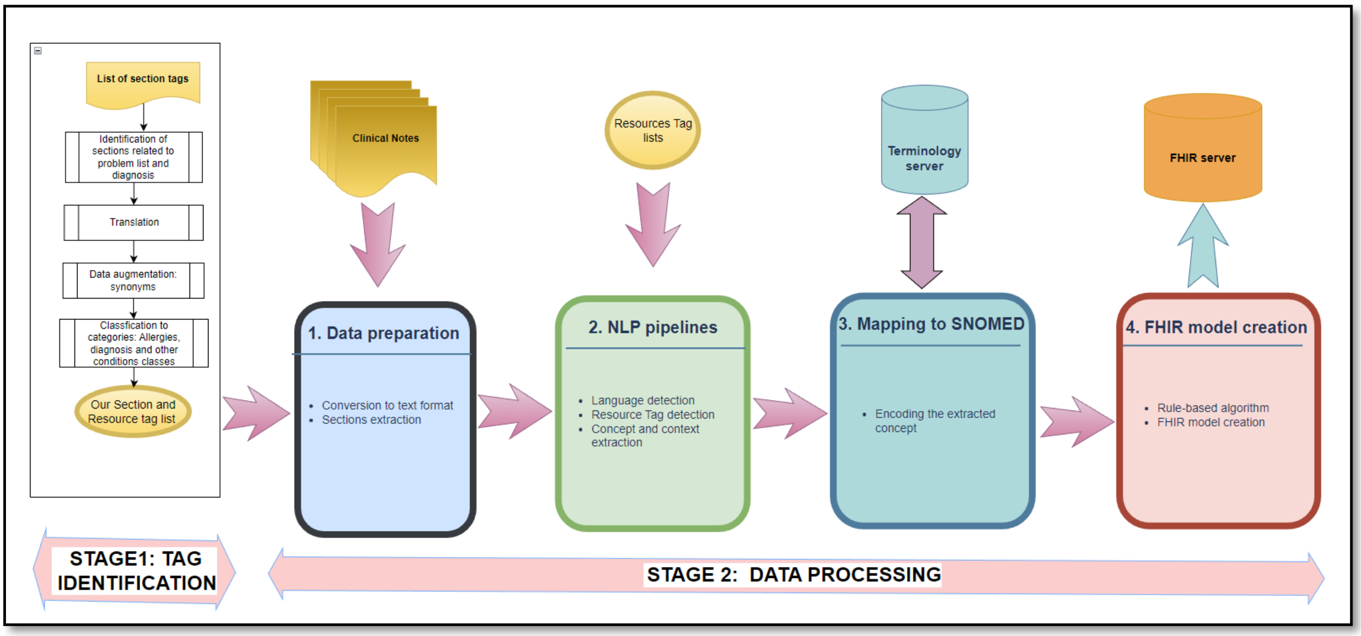

List of section tags
Identification of sections related to problem list and diagnosis
Translation
Data augmentation: synonyms
Classfication to categories: Allergies, diagnosis and other conditions classes
Our Section and Resource tag list
Clinical Notes
Resources Tag lists
Terminology server
FHIR server
1. Data preparation
• Conversion to text format
• Sections extraction
2. NLP pipelines
• Language detection
• Resource Tag detection
• Concept and context extraction
3. Mapping to SNOMED
• Encoding the extracted concept
4. FHIR model creation
• Rule-based algorithm
• FHIR model creation
STAGE1: TAG IDENTIFICATION
STAGE 2: DATA PROCESSING

Figure 1. The framework for converting unstructured clinical problem data using FHIR.

- Stage 2: Data processing

Stage 2 comprises four steps: data preparation, NLP pipelines, mapping to SNOMED, and FHIR model creation (Figure 3).

1. Step 1: Data preparation

The data sources of clinical notes in English or French were converted into text format when required [28]. Next, the list of section tags related to patient problem lists was used to extract the sections where possible.

2. Step 2: NLP pipelines

MedSpacy is an NLP toolkit designed for processing clinical and biomedical texts [29,30]. It is integrated within the SpaCy platform, an open-source NLP library for Python [31].

Language detection: MedSpacy includes tools for language and section detection. Both tools are useful because the clinical notes are written in either English or French. This step is essential because the models are language-dependent, and this study investigated French textual data.

Resource tag detection: This step is implemented with MedSpacy (version 1.3.1) and Python 3.12.3, where the module seeks to detect more context for the concept to determine whether it is related to an allergy or intolerance, as well as any diagnosis, based on the established resource tag list.

Concept and context extraction: To extract clinical concepts with their contexts, the SIFR Annotator was used [32,33]. This is a free French web-based annotating tool that integrates contextual modifiers [34]. It is developed by the laboratory of computer science, robotics and microelectronics of Montpellier in France.

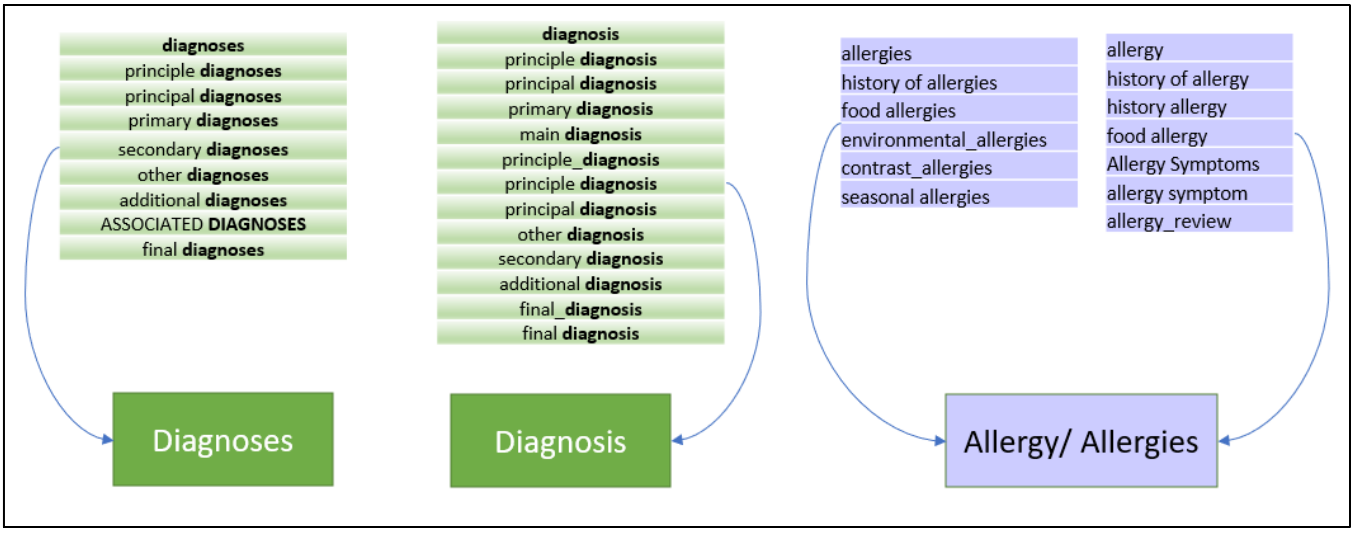


Figure 2. Example of section and resource tag optimization.

3. Step 3: Mapping to SNOMED

Canada Health Infoway has made a cloud-based terminology server available (Ontoserver) to promote semantic interoperability at the Canadian level [35]. Mapping of the extracted concepts to the Canadian SNOMED CT version was performed manually using the Shrimp tool [36].

4. Step 4: FHIR model creation

The data output from step 3 is aggregated to construct the FHIR model. To achieve this, we used the following rule-based method that was designed to map the extracted concepts to their corresponding IPS FHIR elements (Figure 3).

This rule-based method for mapping the NLP output to FHIR model elements consists of

- Four rules related to context modifiers (experiencer, negation, temporality, and certainty);
- Two additional rules related to resource tag identification (allergy, diagnosis).

There are three alternatives for each concept in terms of FHIR resources:

1. Experiencer = non-patient: Experiencer context means that the indicated problem is for the patient or their family member. In this scenario, the FHIR resource used was FamliyMemberHistory. In this study, the focus was on the patient's clinical problems.
2. Experiencer = patient and allergy tag presence: When an allergy tag is detected, the concept is mainly related to an allergy. Therefore, the FHIR resource to be used is

AllergyIntolerance. The negation context is then used to confirm whether the patient had an allergy.

3. Experiencer = patient and no allergy tag detected: The current patient problem is not related to an allergy, so the FHIR resource to be used is Condition. Next, the negation context is used to confirm the decision rule following these alternatives:
    - Negation = Yes (Concept is negated): If it is the only concept for a condition, then the patient has no known conditions, and the corresponding SNOMED CT code will be added. Otherwise, the process continues with the next concept.
    - Negation = No. The next step is to add the corresponding SNOMED CT code and mapping to the corresponding FHIR Condition element following these three rules:
        - ○ Temporality confirms whether the identified problem is still active. The value of the ClinicalStatus element of the Condition resource is active or inactive if the temporality is Recent or Historical, respectively.
        - ○ Certainty enables us to determine whether an identified problem is confirmed or only a hypothesis. In the first case, the VerificationStatus element value is confirmed; otherwise, it is unconfirmed.
        - ○ The diagnosis section tag enables the identification of the condition category element. The possible values are Encounter-diagnosis or Problem-List-Item.

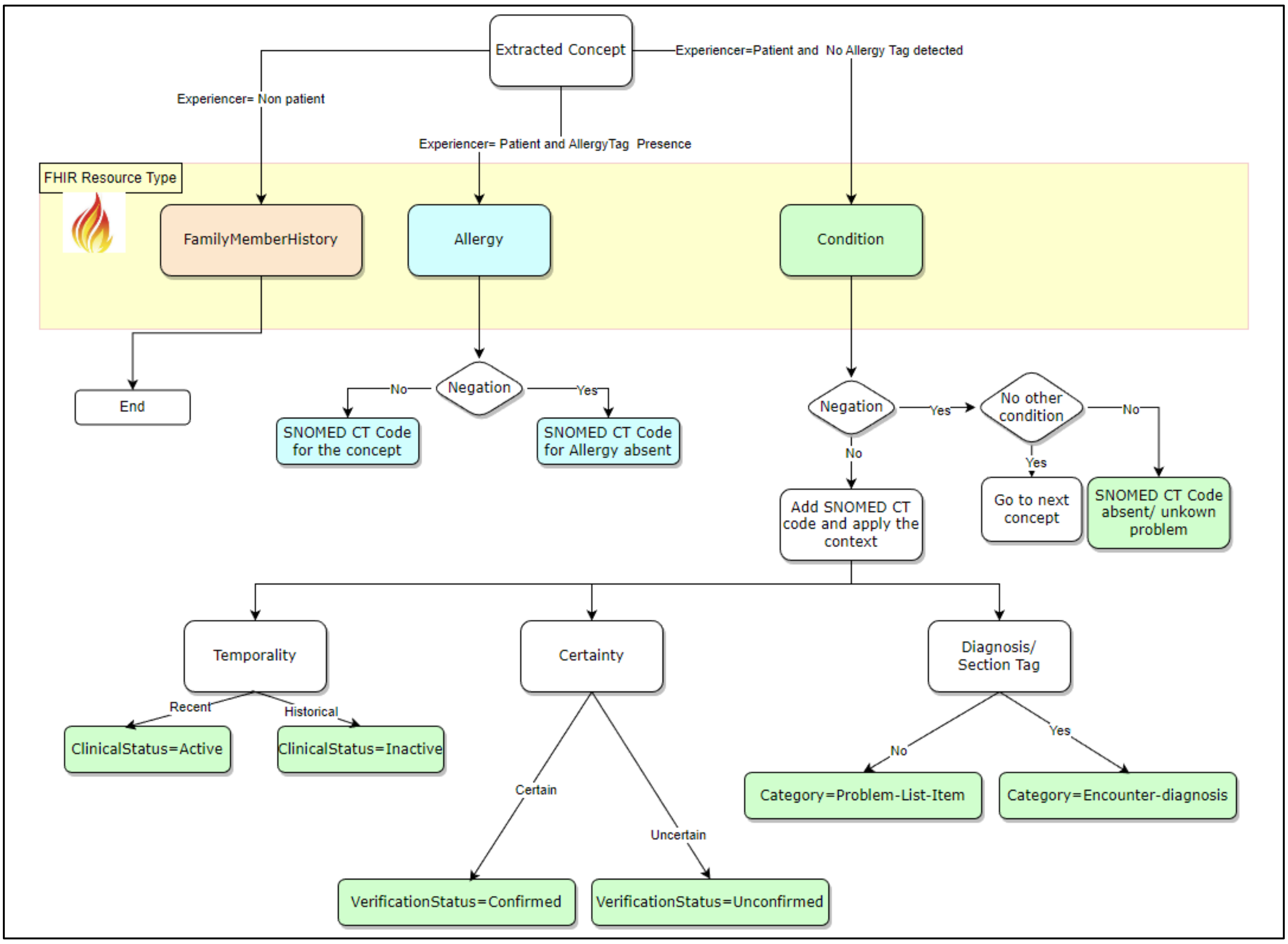

Extracted Concept
Experiencer=Patient and No Allergy Tag detected
Experiencer= Non patient
Experiencer= Patient and AllergyTag Presence
FHIR Resource Type
FamilyMemberHistory
Allergy
Condition
End
No
Negation
Yes
SNOMED CT Code for the concept
SNOMED CT Code for Allergy absent
Negation
Yes
No other condition
No
No
Yes
Add SNOMED CT code and apply the context
Go to next concept
SNOMED CT Code absent/ unkown problem
Temporality
Certainty
Diagnosis/ Section Tag
Recent
Historical
ClinicalStatus=Active
ClinicalStatus=Inactive
Certain
Uncertain
No
Yes
Category=Problem-List-Item
Category=Encounter-diagnosis
VerificationStatus=Confirmed
VerificationStatus=Unconfirmed

Figure 3. Rule-based method for mapping NLP output to FHIR model elements.

3.3. Evaluation

The assessment involves the resource tag detection as well as the rule-based approach:

Resource tag detection: The performance of the implemented module was evaluated using the following standard measures [10–12]: accuracy (Equation (1)), recall (Equation (2)), precision (Equation (3)), and F1 score (Equation (4)).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Equation (1): Accuracy calculation formula

$$R = \frac{TP}{TP + FN} \tag{2}$$

Equation (2): Recall calculation formula

$$P = \frac{TP}{TP + FP} \tag{3}$$

Equation (3): Precision calculation formula

$$F = \frac{2PR}{P + R} \tag{4}$$

Equation (4): F1 score calculation formula

Where TP = true positive, TN = true negative, FP = false positive, and FN = false negative.

- The rule-based approach results were manually validated by testing all possible cases because the dataset did not cover all scenarios of the context modifiers.
- The result of the overall process was then viewed in the IPS viewer.

4. Results

This study used the FRASIMED dataset [5], a French-annotated corpus with 2051 clinical notes. FRASIMED comprises two types of corpora with their corresponding annotated files:

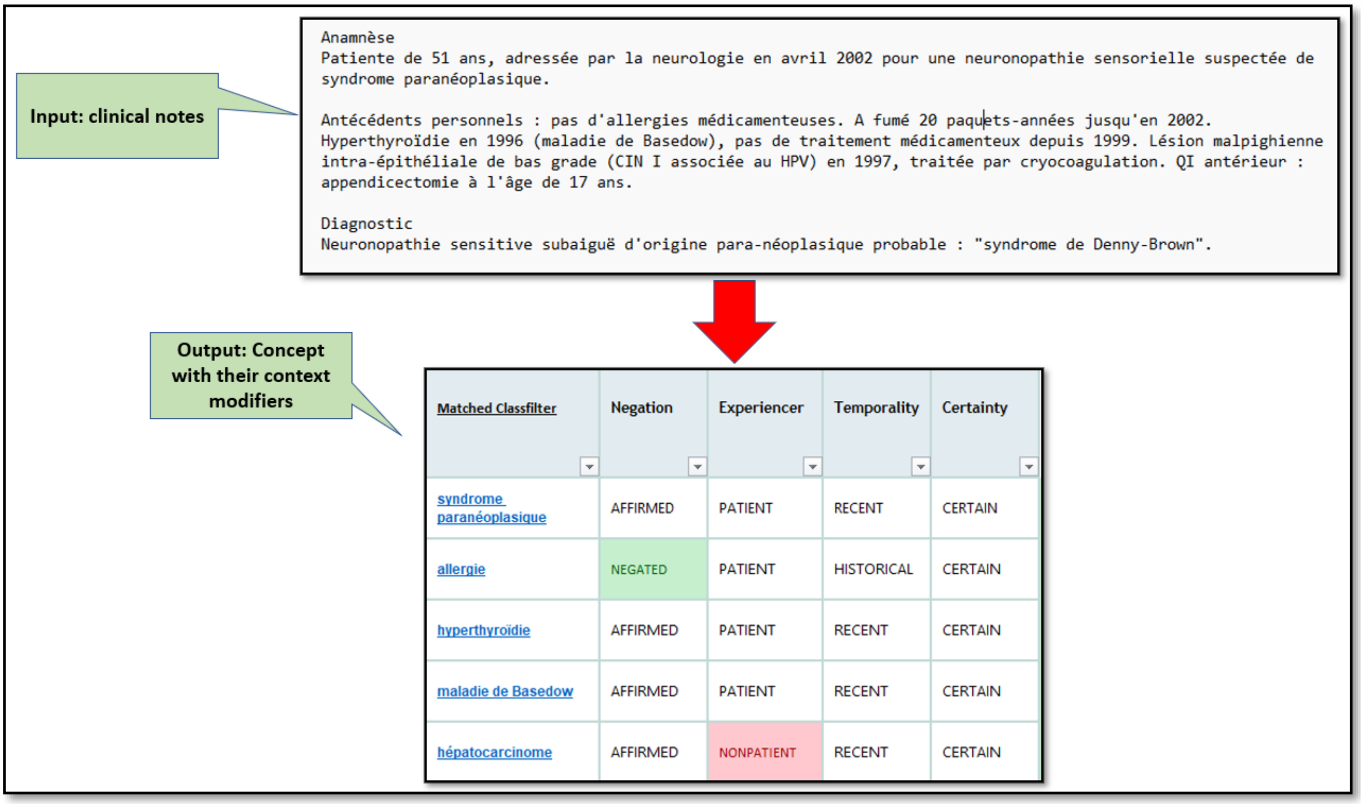
(a) CANTEMIST-FRASIMED: The patient summary is organized into sections for medical history, physical examination, diagnosis, treatment, etc.

(b) DISTEMIST-FRASIMED: The summary is a text with no headers.

A sample of 50 randomly selected clinical notes was used to evaluate the proposed approach (25 files from each FRAMISED corpora).

1. Step 1: Data preparation:
    - Conversion into text format: The file format is text type.
    - Section extraction:
        (a) CANTEMIST-FRASIMED corpus: Based on the resource and section tag list, the sections related to the patient problem list were extracted (see Table 2 for the list of section titles available in this corpus and their corresponding categories);
        (b) DISTEMIST-FRASIMED Corpus: This step is not applicable to this corpus because there are no headers.
2. Step 2: NLP pipelines
    - Language detection: The purpose of this step is to select either the French or English model to be used based on the text file language. This study focuses on French text cases.

- Concepts and context extraction: Figure 4 is an example of the SIFR output for a clinical text that contains the results of context modifiers for each detected concept.



Anamnèse
Patiente de 51 ans, adressée par la neurologie en avril 2002 pour une neuronopathie sensorielle suspectée de syndrome paranéoplasique.

Antécédents personnels : pas d'allergies médicamenteuses. A fumé 20 paquets-années jusqu'en 2002. Hyperthyroïdie en 1996 (maladie de Basedow), pas de traitement médicamenteux depuis 1999. Lésion malpighienne intra-épithéliale de bas grade (CIN I associée au HPV) en 1997, traitée par cryocoagulation. QI antérieur : appendicectomie à l'âge de 17 ans.

Diagnostic
Neuronopathie sensitive subaiguë d'origine para-néoplasique probable : "syndrome de Denny-Brown".

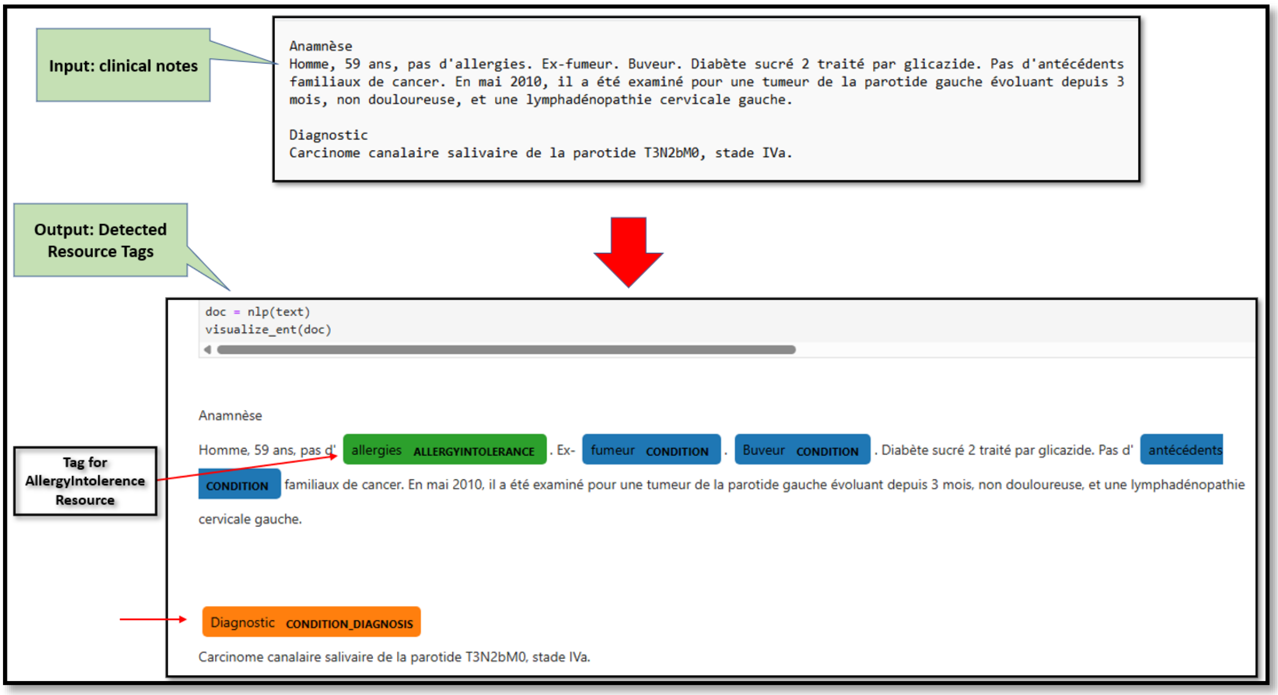
Matched Classfilter	Negation	Experiencer	Temporality	Certainty
syndrome paranéoplasique	AFFIRMED	PATIENT	RECENT	CERTAIN
allergie	NEGATED	PATIENT	HISTORICAL	CERTAIN
hyperthyroïdie	AFFIRMED	PATIENT	RECENT	CERTAIN
maladie de Basedow	AFFIRMED	PATIENT	RECENT	CERTAIN
hépatocarcinome	AFFIRMED	NONPATIENT	RECENT	CERTAIN

Figure 4. An example of the extracted concepts with their context using SIFR.

Table 2. Section selection for the patient problem list in the CANTEMIST_FRAMISED corpus.

Section Title (Original List)	Selected/Unselected	Tag Category
Anamnèse	Selected	Other condition
Examen physique	Unselected	-
Examens complémentaires	Unselected	-
Tests complémentaires	Unselected	-
Diagnostic	Selected	Diagnosis
DIAGNOSTIC PRINCIPAL	Selected	Diagnosis
HISTOIRE DE LA FAMILLE	Selected	Other condition
MALADIE ACTUELLE	Selected	Diagnosis
CONTEXTE PERSONNEL	Selected	Other condition
Antécédents	Selected	Other condition
Antécédents oncologiques	Selected	Other condition
Traitement	Unselected	-
Évolution	Unselected	-
L'évolution	Unselected	-
Développements	Unselected	-

Figure 5 presents an example of the output of our resource tag module for French clinical text. The resource information is highlighted in different colors, and a label is displayed for the corresponding category.


Input: clinical notes
Output: Detected Resource Tags
Tag for AllergyIntolerence Resource

```
Anamnèse
Homme, 59 ans, pas d'allergies. Ex-fumeur. Buveur. Diabète sucré 2 traité par glicazide. Pas d'antécédents
familiaux de cancer. En mai 2010, il a été examiné pour une tumeur de la parotide gauche évoluant depuis 3
mois, non douloureuse, et une lymphadénopathie cervicale gauche.

Diagnostic
Carcinome canalaire salivaire de la parotide T3N2bM0, stade IVa.
```

```
doc = nlp(text)
visualize_ent(doc)
```

Anamnèse

Homme, 59 ans, pas d' allergies ALLERGYINTOLERANCE . Ex- fumeur CONDITION . Buveur CONDITION . Diabète sucré 2 traité par glicazide. Pas d' antécédents CONDITION familiaux de cancer. En mai 2010, il a été examiné pour une tumeur de la parotide gauche évoluant depuis 3 mois, non douloureuse, et une lymphadénopathie cervicale gauche.

Diagnostic CONDITION_DIAGNOSIS

Carcinome canalaire salivaire de la parotide T3N2bM0, stade IVa.

Figure 5. An example of resource tag detection.

The "Allergy resource" tag was evaluated for the French dataset using a sample of 50 randomly selected clinical notes. The evaluation results in terms of accuracy (from 0.947 to 1), recall (1), precision (from 0.8888 to 1), and F1 score (from 0.9411 to 1) are shown in Table 3.

Table 3. Performance evaluation of allergy resource tag on the two FRASIMED files (n = 50).

FRAMISED Dataset Files	Language	Accuracy	Recall	Precision	F1 Score
CANTEMIST-FRASIMED	French	1	1	1	1
DISTEMIST-FRASIMED	French	0.947	1	0.8888	0.9411

The results obtained demonstrate the feasibility of detecting information related to the allergy resource tag from the patient problem list.

3. Step 3: Mapping to SNOMED

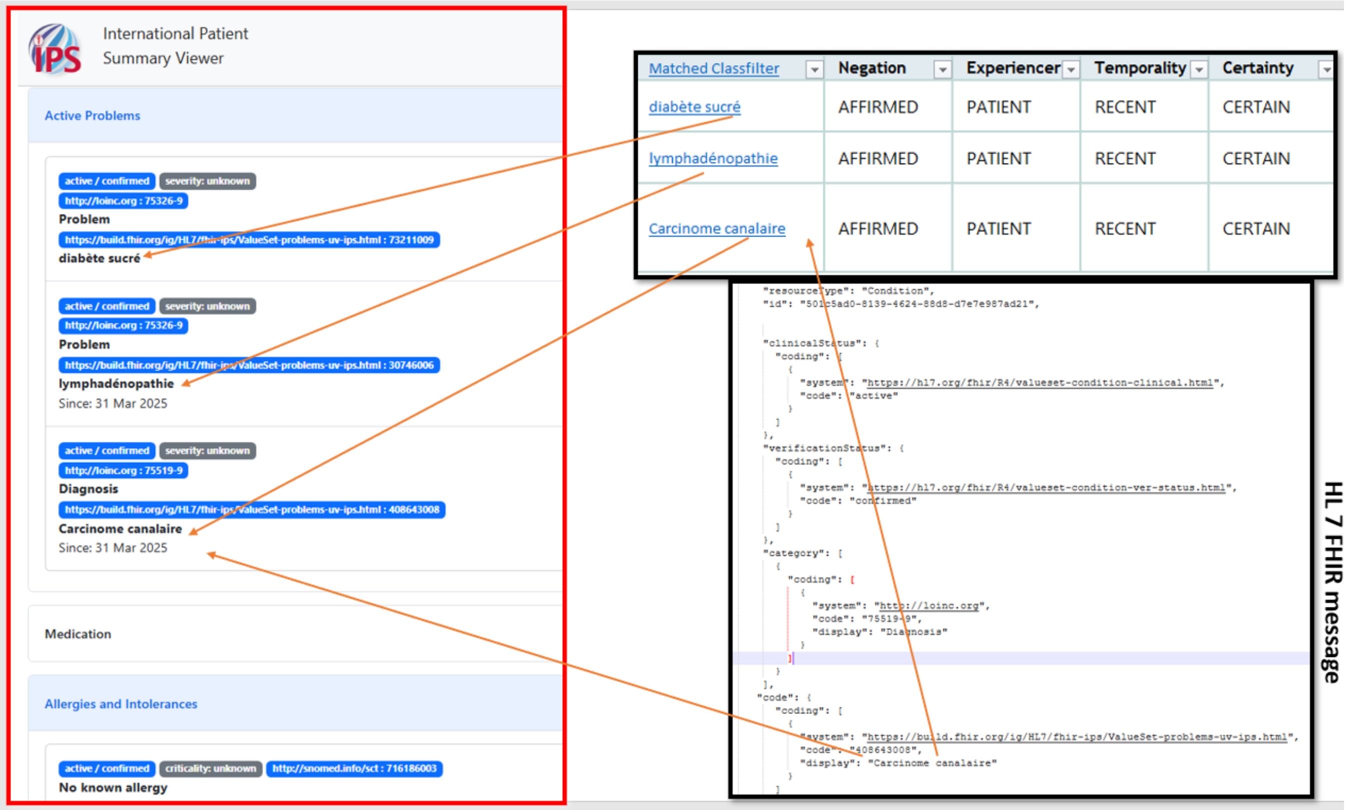
The Shrimp tool was used to find the SNOMED CT code for the extracted concepts related to the problem list.

4. Step 4: FHIR model creation

Before delving deeper into the use cases, it is important to provide a quick overview of the FHIR specifications used.

- Most of the tools used in this study were an implementation of version 4 of FHIR [37].
- The HAPI FHIR server, an implementation of the FHIR specifications in Java, was used to test the proposed FHIR model [38].
- The details of the specification profile describing the FHIR resources and their format are based on the IPS implementation guide [39].

- The IPS viewer is an open-source viewer that allows users to submit an IPS FHIR model and display the information [40]. Figure 6 shows IPS creation using the IPS viewer tool and how context modifiers are considered in the FHIR transaction.



International Patient
Summary Viewer

Active Problems

active / confirmed severity: unknown
http://loinc.org : 75326-9
Problem
https://build.fhir.org/ig/HL7/fhir-ips/ValueSet-problems-uv-ips.html : 73211009
diabète sucré

active / confirmed severity: unknown
http://loinc.org : 75326-9
Problem
https://build.fhir.org/ig/HL7/fhir-ips/ValueSet-problems-uv-ips.html : 30746006
lymphadénopathie
Since: 31 Mar 2025

active / confirmed severity: unknown
http://loinc.org : 75519-9
Diagnosis
https://build.fhir.org/ig/HL7/fhir-ips/ValueSet-problems-uv-ips.html : 408643008
Carcinome canalaire
Since: 31 Mar 2025

Medication

Allergies and Intolerances

active / confirmed criticality: unknown http://snomed.info/sct : 716186003
No known allergy

Matched Classfilter	Negation	Experiencer	Temporality	Certainty
diabète sucré	AFFIRMED	PATIENT	RECENT	CERTAIN
lymphadénopathie	AFFIRMED	PATIENT	RECENT	CERTAIN
Carcinome canalaire	AFFIRMED	PATIENT	RECENT	CERTAIN

```
"resourceType": "Condition",
"id": "501c5ad0-8139-4624-88d8-d7e7e987ad21",


"clinicalStatus": {
  "coding": [
    {
      "system": "https://hl7.org/fhir/R4/valueset-condition-clinical.html",
      "code": "active"
    }
  ]
},
"verificationStatus": {
  "coding": [
    {
      "system": "https://hl7.org/fhir/R4/valueset-condition-ver-status.html",
      "code": "confirmed"
    }
  ]
},
"category": [
  {
    "coding": [
      {
        "system": "http://loinc.org",
        "code": "75519-9",
        "display": "Diagnosis"
      }
    ]
  }
],
"code": {
  "coding": [
    {
      "system": "https://build.fhir.org/ig/HL7/fhir-ips/ValueSet-problems-uv-ips.html",
      "code": "408643008",
      "display": "Carcinome canalaire"
    }
  ]
```

Figure 6. An example of IPS creation using the rule-based approach.

Table 4 describes an example of the concept "Carcinome Canalaire" (Figure 4) (negation = affirmed, experiencer = patient, temporality = recent, certainty = certain).

Table 4. An example of the rule-based validation.

Step	Input	Decision Logic	Output
1	Experiencer = patient and no associated allergy tag	The concept is related to the patient and no flag that it is an allergy	Use the Condition resource
2	Negation	"Affirmed" means there is no negation	Three contexts to verify
3	Temporality	Apply rule for recent	ClinicalStatus = Active
4	Certainty	Apply rule for value = certain	VerificationStatus = Confirmed
5	Diagnostic section tag	Concept included in Diagnosis section	Category = Encounter-diagnosis

5. Limitations

The performance of this framework depends on the NLP output using the annotator, especially for context modifiers. There are still issues regarding abbreviations that are not necessarily detected. Although NLP techniques provide promising results, there are still some limitations, as the clinical notes may include abbreviations, syntax, and grammatical errors that may have a negative impact on the quality of results.

Another important limitation is related to the absence of a gold-standard test for evaluating the full conversion of the patient problem list/condition, including allergies, from unstructured data to well-known terminologies, such as SNOMED CT.

6. Discussion, Contributions, and Future Work

Several studies on clinical data have been devoted to patient summaries, particularly on how to extract and structure valuable information from clinical notes using NLP tools.

This paper presented a method for converting an unformatted patient problem list into a formal model using an IPS profile that can be shared around the globe, satisfying semantic interoperability using SNOMED CT as terminology.

Overall, the framework described in this paper demonstrated the feasibility of converting the patient problem list from French free text into an FHIR-based model using a hybrid system of NLP and a rule-based technique. The proposed approach considers the challenges of context modifiers (negation, experiencer, temporality, and certainty). Two main FHIR resources were used in this study: Condition and Allergy Intolerance.

The primary contributions are the implementation of the module responsible for the early detection of possible FHIR resource presence (resource tag) from unstructured data in the French language (accuracy (from 0.947 to 1), recall (1), precision (from 0.8888 to 1), and F1 score (from 0.9411 to 1), followed by the design of a rule-based algorithm to identify and map the extracted data to the appropriate FHIR resource attributes using an annotator.

Unlike the current models that use translation into English, our framework is based on native French tools, which opens new perspectives to automatically generate and share patient summaries taking into consideration the context modifiers.

Although advanced English-language NLP models and tools like Google Translate offer the capability to process French text, this approach faces the following major challenges in digital translation policy:

- The translation is never perfect, especially in the healthcare domain, as it can lead to semantic loss [41,42].
- The French language is rich and can be finely tuned in a specific context (regulatory, administrative, etc.).
- The development of models in French is crucial to ensure data, digital, and technology sovereignty [43–45]. There is increasingly a need for different forms of independence, control, and autonomy over digital infrastructure, technologies, and data. For example, translating into English for processing potentially exposes data to third parties (US servers). Canadian public health prefers locally hosted solutions due to security, confidentiality, and other regulatory constraints.
- If no one develop NLP models in French, the language will be hidden in AI systems, and furthermore, translating may risk enforcing a linguistic and cultural bias.
- Finally, a well-trained French-language model may outperform an English model when used alongside a translation system.

Although this study focused on the patient problem list, this approach can be applied to other resources as well as the EHR, which contains a minimum amount of structured and semi-structured data.

This research addressed the possibility of converting an unstructured patient problem list into an FHIR model, considering their context modifiers (negation, experiencer, temporality, and certainty). This research focused on French-language texts because there is a significant demand for semantic interoperability in Canada and other French countries around the world.

This paper presented a proposed FHIR-based framework consisting of using an NLP clinical pipeline as well as a rule-based approach to converting the patient problem list, including allergies, into an FHIR model. The proposed approach considers concept modifiers while mapping to the IPS FHIR model elements. The feasibility of the proposed approach is evaluated using the FRASIMED dataset.

Although the approach was limited to the patient problem list in the French language, ongoing work is underway on the interoperability of many other resources, such as immunization, to cover the maximum information available on the clinical patient profile.

Future work will also include the automatization of current manual steps, such as using the terminology server API instead of a manual process. This allows for easier system-to-system access. Furthermore, additional validation and systematic evaluation using real hospital data are underway to enhance the reliability of the performance assessment.

Finally, the proposed framework considered the Canadian context of bilingual clinical notes. This will be useful for other countries with similar national contexts, particularly in Europe and Africa, where both English and French are widely used.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/electronics14214134/s1.

Author Contributions: F.A.: Conceptualization; methodology; formal analysis; software, data curation; writing—original draft; writing—review and editing; visualization. A.A. (Alain April): Writing—review and editing. A.A. (Alain Abran): Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. The dataset used for this work is available as follows: FRASIMED, composed of clinical synthetic cases for French NER and Entity Linking, is freely available, under the Creative Commons 4.0 License, in Zenodo: https://doi.org/10.5281/zenodo.8355629 (accessed on 1 April 2025).

Acknowledgments: During the preparation of this study, the authors used ChatGPT for the tag synonyms, as indicated in the manuscript. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
BioBERT	Bidirectional Encoder Representations from Transformers for Biomedical Text Mining
CEN	European Committee for Standardization
CDA	Clinical Document Architecture
CLAMP	Clinical Language Annotation, Modeling, and Processing
cTAKES	Clinical Text Analysis and Knowledge Extraction System
eHN	European eHealth Network
EHR	Electronic Health Record
FHIR	Fast Healthcare Interoperability Resource
G7	Group of Seven Summits: Annual meeting of leaders from seven of the world's largest advanced economies
GDHP	Global Digital Health Partnership
HL7	Health Level Seven
IHE	Integrated Healthcare Exchange
IPS	International Patient Summary
ISO	International Organization for Standardization
PS-CA	Canadian Patient Summary
MedXN	Medication eXtraction and Normalization
MedSpacy	SpaCy-based library of core components targeting medical text
ML	Machine Learning

NLM	National Library of Medicine
NLP	Natural Language Processing
ONC	Office of the National Coordinator
SIFR	Ontology-based annotation web service to process biomedical text in French
SNOMED CT	Systematized Nomenclature of Medicine—Clinical Terms
UIMA	Unstructured Information Management Architecture
UMLS	Unified Medical language System

References

1. International Patient Summary. 2025. Available online: https://international-patient-summary.net/ips-links-to-standards-and-specifications/ (accessed on 1 April 2025).
2. Amar, F.; April, A.; Abran, A. Electronic Health Record and Semantic Issues Using Fast Healthcare Interoperability Resources: Systematic Mapping Review. *J. Med. Internet Res.* **2024**, *26*, e45209. [CrossRef]
3. Health Level Seven International IPS. 2025. Available online: https://hl7.org/fhir/uv/ips/ (accessed on 1 April 2025).
4. HL7, IPS-Condition Resource. 2025. Available online: https://build.fhir.org/ig/HL7/fhir-ips/StructureDefinition-Condition-uv-ips.html (accessed on 1 April 2025).
5. Zaghir, J.; Bjelogrlic, M.; Goldman, J.-P.; Aananou, S.; Gaudet-Blavignac, C.; Lovis, C. FRASIMED: A Clinical French Annotated Resource Produced through Crosslingual BERT-Based Annotation Projection. *arXiv* **2023**, arXiv:2309.10770. [CrossRef]
6. Durango, M.C.; Torres-Silva, E.A.; Orozco-Duque, A. Named Entity Recognition in Electronic Health Records: A Methodological Review. *Healthc. Inform. Res.* **2023**, *29*, 286–300. [CrossRef]
7. Gaudet-Blavignac, C.; Foufi, V.; Bjelogrlic, M.; Lovis, C. Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review. *J. Med. Internet Res.* **2021**, *23*, e24594. [CrossRef] [PubMed]
8. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
9. Hong, N.; Wen, A.; Stone, D.J.; Tsuji, S.; Kingsbury, P.R.; Rasmussen, L.V.; Pacheco, J.A.; Adekkanattu, P.; Wang, F.; Luo, Y.; et al. Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. *J. Biomed. Inform.* **2019**, *99*, 103310. [CrossRef]
10. Hong, N.; Wen, A.; Shen, F.; Sohn, S.; Liu, S.; Liu, H.; Jiang, G. Integrating Structured and Unstructured EHR Data Using an FHIR-based Type System: A Case Study with Medication Data. *Proc. AMIA Jt. Summits Transl. Sci.* **2018**, *2017*, 74–83.
11. Hong, N.; Wen, A.; Shen, F.; Sohn, S.; Wang, C.; Liu, H.; Jiang, G. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. *JAMIA Open* **2019**, *2*, 570–579. [CrossRef]
12. Liu, S.; Luo, Y.; Stone, D.; Zong, N.; Wen, A.; Yu, Y.; Rasmussen, L.V.; Wang, F.; Pathak, J.; Liu, H.; et al. Integration of NLP2FHIR Representation with Deep Learning Models for EHR Phenotyping: A Pilot Study on Obesity Datasets. *AMIA Jt. Summits Transl. Sci. Proc.* **2021**, *2021*, 410–419. [PubMed]
13. Soysal, E.; Wang, J.; Jiang, M.; Wu, Y.; Pakhomov, S.; Liu, H.; Xu, H. CLAMP—A toolkit for efficiently building customized clinical natural language processing pipelines. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 331–336. [CrossRef]
14. Wang, J.; Mathews, W.C.; Pham, H.A.; Xu, H.; Zhang, Y. Opioid2FHIR: A system for extracting FHIR-compatible opioid prescriptions from clinical text. In Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Republic of Korea, 16–19 December 2020; pp. 1748–1751. [CrossRef]
15. Peterson, K.J.; Jiang, G.; Liu, H. A corpus-driven standardization framework for encoding clinical problems with HL7 FHIR. *J. Biomed. Inform.* **2020**, *110*, 103541. [CrossRef]
16. Peterson, K.J.; Liu, H. Automating the Transformation of Free-Text Clinical Problems into SNOMED CT Expressions. *AMIA Jt. Summits Transl. Sci. Proc.* **2020**, *2020*, 497–506.
17. Wu, H.; Toti, G.; Morley, K.I.; Ibrahim, Z.M.; Folarin, A.; Jackson, R.; Kartoglu, I.; Agrawal, A.; Stringer, C.; Gale, D.; et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 530–537. [CrossRef]
18. SNOMED CT Tooling. 2025. Available online: https://www.snomed.org/software-tools (accessed on 1 March 2025).
19. Shaitarova, A.; Zaghir, J.; Lavelli, A.; Krauthammer, M.; Rinaldi, F. Exploring the Latest Highlights in Medical Natural Language Processing across Multiple Languages: A Survey. *Yearb. Med. Inform.* **2023**, *32*, 230–243. [CrossRef]
20. Denny, J.C.; Spickard, A.; Johnson, K.B.; Peterson, N.B.; Peterson, J.F.; Miller, R.A. Evaluation of a Method to Identify and Categorize Section Headers in Clinical Documents. *J. Am. Med. Inform. Assoc.* **2009**, *16*, 806–815. [CrossRef] [PubMed]

21. Pomares-Quimbaya, A.; Kreuzthaler, M.; Schulz, S. Current approaches to identify sections within clinical narratives from electronic health records: A systematic review. *BMC Med. Res. Methodol.* **2019**, *19*, 155. [CrossRef]
22. Deepl Translator. 2025. Available online: https://www.deepl.com/en/translator (accessed on 1 March 2025).
23. French English Medical Dictionary. 2025. Available online: https://dictionary.reverso.net/medical-french-english/ (accessed on 1 March 2025).
24. ChatGPT. 2025. Available online: https://chatgpt.com/ (accessed on 1 March 2025).
25. Feng, S.Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; Hovy, E. A Survey of Data Augmentation Approaches for NLP. *arXiv* **2021**, arXiv:2105.03075. [CrossRef]
26. Bayer, M.; Kaufhold, M.-A.; Reuter, C. A Survey on Data Augmentation for Text Classification. *ACM Comput. Surv.* **2023**, *55*, 1–39. [CrossRef]
27. Li, B.; Hou, Y.; Che, W. Data augmentation approaches in natural language processing: A survey. *AI Open* **2022**, *3*, 71–90. [CrossRef]
28. File Convertor PDF to Text. 2025. Available online: https://www.freeconvert.com/pdf-to-text (accessed on 1 November 2024).
29. MedspaCy. 2025. Available online: https://github.com/medspacy/medspacy/blob/master/README.md (accessed on 1 November 2024).
30. Eyre, H.; Chapman, A.B.; Peterson, K.S.; Shi, J.; Alba, P.R.; Jones, M.M.; Box, T.L.; DuVall, S.L.; Patterson, O.V. Launching into clinical space with medspaCy: A new clinical text processing toolkit in Python. *arXiv* **2021**, arXiv:2106.07799. [CrossRef]
31. Spacy. 2025. Available online: https://spacy.io/ (accessed on 1 November 2024).
32. SIFR, Clinical French Annotator. 2025. Available online: https://bioportal.lirmm.fr/annotator (accessed on 1 April 2025).
33. Tchechmedjiev, A.; Abdaoui, A.; Emonet, V.; Zevio, S.; Jonquet, C. SIFR annotator: Ontology-based semantic annotation of French biomedical text and clinical notes. *BMC Bioinform.* **2018**, *19*, 405. [CrossRef]
34. Mirzapour, M.; Abdaoui, A.; Tchechmedjiev, A.; Digan, W.; Bringay, S.; Jonquet, C. French FastContext: A publicly accessible system for detecting negation, temporality and experiencer in French clinical notes. *J. Biomed. Inform.* **2021**, *117*, 103733. [CrossRef] [PubMed]
35. Canada Health Infoway Terminology Server. 2025. Available online: https://infocentral.infoway-inforoute.ca/en/tools/standards-tools/terminology-server (accessed on 1 April 2025).
36. Shrimp Tool. 2025. Available online: https://ontoserver.csiro.au/shrimp/ (accessed on 1 April 2025).
37. H7 FHIR V4. 2025. Available online: https://hl7.org/fhir/R4/resourcelist.html (accessed on 1 November 2024).
38. HAPI FHIR. 2025. Available online: https://hapi.fhir.org/ (accessed on 1 November 2024).
39. HL:7, IPS Implementation GuideHL:7, IPS Implementation Guide. 2025. Available online: https://build.fhir.org/ig/HL7/fhir-ips/OperationDefinition-summary.html (accessed on 1 April 2025).
40. Osornio, A.L.; Kaminker, D.; Campos, F.; D'Amore, J. IPS Viewer. 2025. Available online: https://www.ipsviewer.com/classic (accessed on 1 April 2025).
41. Kong, M.; Fernandez, A.; Bains, J.; Milisavljevic, A.; Brooks, K.C.; Shanmugam, A.; Avilez, L.; Li, J.; Honcharov, V.; Yang, A.; et al. Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: A comparative analysis. *BMJ Qual. Saf.* **2025**. *online ahead of print*. [CrossRef]
42. Brandenberger, J.; Stedman, I.; Stancati, N.; Sappleton, K.; Kanathasan, S.; Fayyaz, J.; Singh, D. Using artificial intelligence based language interpretation in non-urgent paediatric emergency consultations: A clinical performance test and legal evaluation. *BMC Health Serv. Res.* **2025**, *25*, 138. [CrossRef] [PubMed]
43. Pizzul, D.; Veneziano, M. Digital sovereignty or sovereignism? Investigating the political discourse on digital contact tracing apps in France. *Inf. Commun. Soc.* **2024**, *27*, 1008–1024. [CrossRef]
44. Tan, K.L.; Chi, C.-H.; Lam, K.-Y. Survey on Digital Sovereignty and Identity: From Digitization to Digitalization. *ACM Comput. Surv.* **2024**, *56*, 1–36. [CrossRef]
45. Couture, S.; Toupin, S. What does the notion of "sovereignty" mean when referring to the digital? *New Media Soc.* **2019**, *21*, 2305–2322. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.