Construction neuro-symbolique d'une ontologie clinique à partir de dossiers patients non structurés : application au sepsis

Eya BEN DHIAF[1], Yasmina SADI[2], Lamjed BEN SAID[1],
Nicholas HEMING[3], Djilali ANNANE[3,4], Zaineb CHELLY DAGDIA[5,6]

[1] Laboratoire Strategies for Modelling and ARtificial inTelligence , SMARTLAB
[2] Laboratoire Informatique, BioInformatique, Systèmes Complexes , IBISC
[3] Service de médecine intensive-réanimation Hôpital Raymond-Poincaré
[4] IHU SEPSIS, U1173 Lab 2I, School of Medicine Simone Veil,
University Versailles SQY, University Paris Saclay
[5] Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL
[6] Univ. Lille, Inserm, CHU Lille, U1286, INFINITE, Institute for Translational Research in Inflammation

11 mai 2026

Résumé

La détection précoce du sepsis demeure un défi clinique majeur en raison de la variabilité symptomatique, de la fragmentation des données médicales et de la prédominance des textes non structurés dans les dossiers patients. Malgré les progrès de l'intelligence artificielle, les approches purement neuronales manquent d'interprétabilité, tandis que les systèmes symboliques traditionnels peinent à intégrer des données hétérogènes. Faisant partie des projets de l'IHU SEPSIS[1], ce travail propose KG-EDS (Knowledge Graph for Early Diagnosis of Sepsis), un cadre neuro-symbolique dont les deux contributions principales sont : (1) une stratégie d'extraction d'informations cliniques par apprentissage en contexte (In-Context Learning, ICL) sans corpus spécialisé disponible pour le fine-tuning, et (2) la construction d'une ontologie clinique orientée diagnostic précoce du sepsis, validée par raisonnement hybride sur graphe de connaissances. Notre approche est évaluée sur un échantillon de 58 dossiers de la cohorte APROCCHSS (1 240 patients), fournie par l'AP-HP, France. L'ontologie produite comprend 363 classes et 380 propriétés d'objets, validées par questions de compétence et vérification de cohérence logique. Cette étude sera complétée par l'application du pipeline proposé sur tous les dossiers de la cohorte et la validation de l'ontologie par les professionnels de santé.

Mots-clés

Intelligence artificielle neuro-symbolique ; Ontologie clinique ; Graphe de connaissances ; Aide à la décision clinique ; Dossiers de santé électroniques ; Sepsis ; Intelligence artificielle explicable ; Grands modèles de langage ; Reconnaissance optique de caractères ; Apprentissage en contexte ; Reconnaissance d'entités nommées ; Extraction de relations.

Abstract

Early detection of sepsis remains a major clinical challenge due to symptom variability, fragmentation of medical data, and the predominance of unstructured text in patient records. Despite advances in artificial intelligence, purely neural approaches lack interpretability, while traditional symbolic systems struggle to integrate heterogeneous data. As part of the projects of the Institut Hospitalo-Universitaire (IHU) sepsis[1], this work proposes KG-EDS (Knowledge Graph for Early Diagnosis of Sepsis), a neuro-symbolic framework whose two main contributions are : (1) a clinical information extraction strategy based on In-Context Learning (ICL) without an annotated corpus, and (2) the construction of a clinically oriented ontology for early diagnosis of sepsis, automatically populated and validated through hybrid reasoning on a knowledge graph. Our approach is evaluated on a sample of 58 records from the APROCCHSS cohort (1,240 patients), provided by AP-HP, France. The resulting ontology comprises 363 classes and 380 object properties, validated through competency questions and logical consistency checking. This study will be completed by applying the proposed pipeline to all records in the cohort and by validating the ontology with healthcare professionals.

Keywords

Neuro-symbolic AI ; Clinical ontology ; Knowledge graph ; Clinical Decision Support System ; Electronic Health Records ; Sepsis ; Explainable AI ; Large Language Models ; Optical Character Recognition ; In-Context Learning ; Named Entity Recognition ; Relation Extraction.

1 Introduction

Sur le plan clinique, le sepsis résulte d'une réponse immunitaire dérégulée à l'infection conduisant à une défaillance multi-organique, dont la prise en charge repose sur l'anti-

1. ANR-23-IAHU-0004

biothérapie précoce, la réanimation hémodynamique et un soutien organique adapté, dans le cadre des recommandations de la Surviving Sepsis Campaign [25]. Le sepsis représente une cause majeure de mortalité mondiale. L'analyse GBD (Global Burden of Diseases) 2021 [10] estime à 166 millions le nombre de cas et à 21,4 millions le nombre de décès en 2021, soit 31,5% de la mortalité mondiale toutes causes confondues, avec une recrudescence liée à la pandémie de COVID-19 après une période de déclin entre 1990 et 2019. La charge est croissante chez les personnes de plus de 15 ans et s'observe de façon préoccupante en complication de maladies chroniques non infectieuses telles que l'AVC, la BPCO et le diabète. Des inégalités géographiques majeures persistent, avec une charge disproportionnée dans les pays à ressources limitées. Malgré les progrès réalisés en soins intensifs et en thérapies antimicrobiennes, le diagnostic précoce demeure complexe en raison de l'hétérogénéité et du caractère non spécifique des manifestations cliniques. Un retard de reconnaissance augmente significativement le risque de mortalité, en particulier dans les services d'urgence et de soins critiques où la rapidité décisionnelle est déterminante.

Les approches diagnostiques traditionnelles reposent sur des systèmes de scores fondés sur des règles, tels que SIRS, qSOFA et SOFA. Bien que le SIRS présente une sensibilité élevée, il manque de spécificité et peut entraîner un nombre important de faux positifs [20, 23, 5]. À l'inverse, le qSOFA offre une meilleure spécificité mais tend à identifier le sepsis à des stades plus avancés [5, 21]. Le score SOFA, au cœur de la définition sepsis-3 [24], possède une valeur pronostique plus robuste, mais dépend de paramètres biologiques qui ne sont pas toujours disponibles immédiatement [5, 21]. Ces limites soulignent la nécessité d'approches computationnelles plus adaptatives et sensibles au contexte clinique.

Les méthodes d'apprentissage automatique (ML) et d'apprentissage profond (DL) se sont imposées comme des solutions prometteuses pour la prédiction précoce du sepsis à partir des dossiers patients informatisés (DPI) [14]. Les modèles classiques tels que Random Forest et Gradient Boosting affichent des performances compétitives [14], tandis que les architectures séquentielles comme les LSTM (Long Short-Term Memory) permettent de capturer la dynamique temporelle des données physiologiques. L'intégration de variables structurées et de textes cliniques non structurés améliore encore la précision prédictive [29, 1]. Toutefois, de nombreux modèles existants se concentrent davantage sur le pronostic que sur le soutien au diagnostic précoce [14], et leur nature de « boîte noire » soulève des préoccupations en matière d'interprétabilité et de confiance clinique.

Les techniques de traitement automatique du langage naturel (NLP) ont considérablement progressé dans l'extraction d'informations (Information Extraction, IE) à partir des textes cliniques, facilitant la reconnaissance d'entités et l'extraction de relations. Néanmoins, elles négligent souvent la cohérence sémantique et les connaissances spécifiques au domaine médical. Les ontologies biomédicales et les graphes de connaissances (Knowledge Graphs, KG) apparaissent comme des outils prometteurs pour répondre à ces enjeux, en fournissant des représentations structurées reliant entités (par exemple, symptômes, biomarqueurs) et relations, favorisant ainsi le raisonnement et l'interopérabilité [11]. Cependant, les cadres existants combinent rarement la numérisation robuste des textes, l'extraction avancée d'informations et le raisonnement guidé par ontologie au sein d'un pipeline unifié, notamment pour des DPI hétérogènes, numérisés et rédigés en langues autres que l'anglais, comme le français. Cette lacune limite la transparence, l'adaptabilité et l'utilité clinique des systèmes d'aide à la décision, en particulier dans des contextes à ressources limitées où la fragmentation des données est marquée.

Afin de répondre aux problématiques de fragmentation des données et de manque d'interprétabilité, les graphes de connaissances pilotés par ontologie offrent un cadre sémantique pour intégrer des informations cliniques hétérogènes [3, 11, 15]. Des travaux récents démontrent le potentiel des approches fondées sur les graphes pour améliorer la détection précoce du sepsis et renforcer l'explicabilité des modèles [17, 30].

Nous proposons KG-EDS, un cadre neuro-symbolique complet visant à transformer des dossiers cliniques français non structurés en un graphe de connaissances orienté décision. Nos deux contributions principales sont développées dans cet article :

1. **Une stratégie d'extraction d'informations cliniques par apprentissage en contexte sans fine-tuning supervisé** (section 4), reposant sur l'apprentissage en contexte (In Context Learning, ICL) contraint par schéma terminologique. Nous démontrons que des LLM génériques peuvent produire des sorties structurées exploitables sans fine-tuning ni corpus annoté de référence, réduisant ainsi les coûts d'annotation tout en maintenant une cohérence sémantique contrôlée et nous justifions notre choix par rapport aux méthodes supervisées classiques.
2. **La construction d'une ontologie clinique orientée diagnostic précoce et son intégration dans un graphe de connaissances raisonnant** (section 5), comprenant l'alignement terminologique sur six terminologies biomédicales de référence, la modélisation en triplets RDF et un raisonnement hybride multicouches pour le support à la décision clinique.

Un pipeline OCR hybride (Donut Transformer + Tesseract) constitue le prérequis technique permettant de traiter les DPI numérisés bruités. Il est décrit brièvement en section 3.

2 Travaux Connexes

ICD [27], SNOMED CT [8], MeSH [26] et LOINC [22] sont des exemples de terminologies biomédicales de référence facilitant l'interopérabilité sémantique, respectivement pour la classification des maladies, la structuration des données cliniques, l'indexation bibliographique et l'échange des résultats de laboratoire.

Les ontologies peuvent être utilisées pour fournir des alertes de prescription en temps réel et garantir la sécurité médicamenteuse, comme le démontrent OntoPharma [6] et d'autres systèmes d'aide à la décision clinique pilotés par ontologie. Les graphes de connaissances, associés à une ontologie domaine, permettent de représenter des données hétérogènes pour soutenir le raisonnement pronostique et thérapeutique. C'est l'approche adoptée pour prédire l'évolution de la leucémie lymphoïde chronique avec l'ontologie CLLPO (Chronic Lymphocytic Leukemia Prognosis Ontology) [19], dont les concepts ont été alignés avec des ontologies existantes telles que BO (Blood Ontology) et CLO (Cancer Cell Ontology). D'autres initiatives, telle que CIDO [13], illustrent comment les maladies infectieuses et l'annotation basée sur ontologie peuvent être intégrées pour soutenir la recherche sur la COVID-19 [5, 9], le repositionnement de médicaments et le développement de vaccins.

2.1 Ontologies et graphes de connaissances pour l'aide à la décision clinique dans le diadnostic du sepsis

Les systèmes basés sur des ontologies et les graphes de connaissances (KG) se révèlent être des solutions prometteuses pour améliorer l'interprétabilité et le raisonnement sémantique. À partir de plus de 300 000 résumés PubMed, des recherches antérieures ont développé des approches d'apprentissage ontologique pour le sepsis. Pour extraire des termes biomédicaux et les mapper sur des concepts UMLS au format OWL, le système POLEOS utilisait des modèles de deep learning (CBOW et Skip-gram). Les modèles neuronaux surpassaient les techniques statistiques classiques (LSA, LDA) en termes de précision (>85%), démontrant l'efficacité des embeddings profonds pour l'identification automatique de concepts biomédicaux liés au sepsis [2]. Des ontologies spécialisées, générées manuellement ou automatiquement et ciblant les connaissances sur le sepsis, incluent SEPON [4] et POLEOS [2].

Une ontologie combinant AAENOTE et CIIO [28] a été proposée pour représenter les événements indésirables associés au sepsis. À partir de textes cliniques annotés, elle reliait patients, dispositifs et indicateurs d'infection. La double structure AAENOTE pour les annotations et CIIO pour le raisonnement permet à la fois un étiquetage explicite des données et un raisonnement implicite (par exemple, l'identification des infections liées aux cathéters). Cette approche illustre comment une conception ontologique guidée par les cliniciens améliore l'explicabilité et l'utilisabilité en contexte clinique [28].

Les graphes de connaissances permettent de modéliser de manière structurée les relations entre symptômes, biomarqueurs, diagnostics et paramètres physiologiques [15]. En intégrant des données cliniques hétérogènes dans des représentations sémantiques unifiées, les graphes de connaissances (Knowledge Graph, KG) facilitent le raisonnement contextuel et l'inférence transparente. OntoSepsisKG [7] est une méthodologie de construction d'un graphe de connaissances fondée sur des données transcriptomiques provenant de patients atteints de sepsis. Cette approche s'appuie sur Gene Ontology comme ontologie de domaine et intègre les voies biologiques issues de Reactome afin de représenter les relations entre gènes, protéines et processus biologiques impliqués dans la maladie. Le graphe met en évidence des mécanismes associés à la dérégulation immunitaire et à la dysfonction mitochondriale observées dans le sepsis. Il permet également de relier les données d'expression génique à des connaissances biologiques structurées, facilitant ainsi l'interprétation biologique des interactions moléculaires et des voies biologiques impliquées dans les mécanismes moléculaires du sepsis. Un cadre assisté par LLM pour la création d'un graphe de connaissances sur le sepsis, utilisant GPT-4 et des ontologies médicales (UMLS, SNOMED-CT, ICD), a été présenté récemment. Le système organise les notes cliniques pour le raisonnement en extrayant automatiquement entités et relations. Cette approche neuro-symbolique favorise l'IA explicable dans la prise en charge du sepsis, en combinant l'interprétabilité des ontologies avec la compréhension contextuelle des LLM [30]. Les KG pilotés par ontologie offrent robustesse, explicabilité, interopérabilité et alignement avec les terminologies médicales standardisées [12, 16, 15]. Néanmoins, des défis persistent concernant l'hétérogénéité des jeux de données, l'incohérence des définitions du sepsis et la généralisation limitée.

Alors que ces travaux ont jeté les bases de la structuration sémantique, les premières études se concentraient principalement sur l'apprentissage ontologique à partir de la littérature biomédicale et offraient peu d'intégration clinique [2]. Les travaux ultérieurs, tels que l'ontologie SEPON [4] et ceux de [28], ont amélioré la cohérence sémantique et l'alignement terminologique, mais n'ont pas encore abouti à la construction de graphes de connaissances ou au raisonnement automatisé. Les efforts actuels, comme dans [30] et [17], explorent des stratégies hybrides combinant raisonnement piloté par ontologie et LLM, mais restent limités en termes de reproductibilité, d'évolutivité et de modélisation temporelle.

3 Aperçu du système proposé KG-EDS

Le cadre vise à combler le décalage entre les textes médicaux non structurés et le raisonnement clinique structuré en combinant trois paradigmes complémentaires : compréhension de documents basée sur la vision, extraction d'informations basée sur le langage et représentation des connaissances pilotée par ontologie. KG-EDS met ainsi en œuvre une approche unifiée transformant les dossiers médicaux numérisés et textuels en structures de connaissances interprétables, interrogeables et sémantiquement cohérentes.

L'architecture est organisée en cinq étapes, illustrées à la Figure 1, chacune correspondant à une couche fonctionnelle du pipeline :

1. **Prétraitement et préparation des données** – pipeline hybride de numérisation OCR intégrant les moteurs Donut Transformer et Tesseract, optimisé par recherche aléatoire, pour extraire un texte de haute

qualité à partir de DPI bruités. Donut améliore la robustesse contextuelle et la compréhension de la mise en page, tandis que Tesseract assure efficacité et support multilingue.

2. **Extraction d'informations avec des LLM** – module basé sur l'In-Context Learning (ICL) pour la Reconnaissance d'Entités Nommées (Named Entity Recognition, NER) et l'Extraction de Relations (Relation Extraction, RE) (section 4).
3. **Construction de l'ontologie et du graphe de connaissances** – conception ontologique pour construire un graphe de connaissances sur le sepsis cohérent et interopérable.
4. **Raisonnement et inférence** – couche décisionnelle hybride combinant logique basée sur ontologie, récupération de similarité via embeddings et raisonnement contextuel piloté par LLM.
5. **Intégration dans les systèmes d'aide à la décision clinique** – couche finale intégrant le raisonnement ontologique et les mécanismes d'inférence dans les workflows cliniques, permettant des alertes explicables et un soutien au diagnostic.

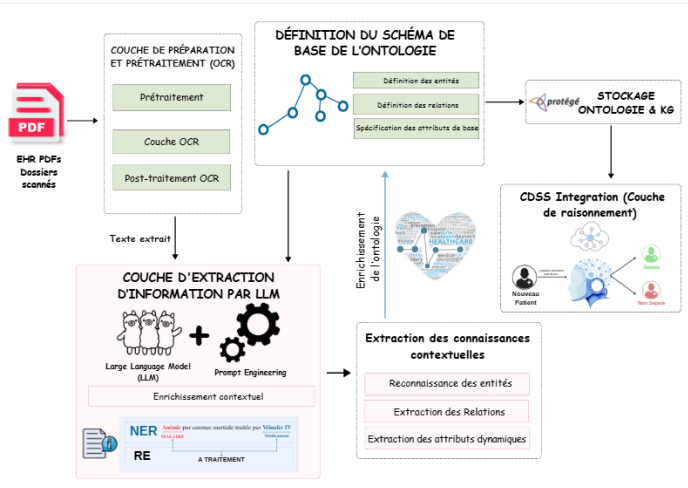

FIGURE 1 – Vue d'ensemble du système KG-EDS

4 Extraction d'informations cliniques par apprentissage en contexte

Un défi majeur en NLP clinique, notamment pour la reconnaissance des entités nommées et l'extraction des relations, est la rareté des corpus annotés, en particulier pour des textes médicaux en français. Il existe deux grandes familles d'approches pour l'extraction d'entités nommées et de relations cliniques :

1. **Les approches supervisées :** modèles de type BERT spécialisés comme BioBERT, ClinicalBERT OU DrBERT pour le français mais nécessitent un corpus annoté spécifique au domaine cible pour le fine-tuning. A notre connaissance, il n'existe pas actuellement de corpus clinique francophone spécifiquement pour le sepsis et directement exploitable pour un fine-tuning supervisée.
2. **Les approches par apprentissage en contexte (In-Context Learning, ICL) :** elles exploitent les capacités de généralisation des LLM pour produire des sorties structurées à partir d'exemples intégrés au prompt, sans ajustement de paramètres du modèle.

Dans ce contexte, nous avons privilégié une stratégie d'apprentissage en contexte (ICL) reposant sur la construction manuelle d'exemples few-shot intégrés aux prompts. L'utilisation de modèles comme Mistral-Small 3.2 permet d'exploiter une robustesse multilingue adaptée aux spécificités du français médical.

Prompt Engineering et adaptation des LLM pour l'extraction d'informations cliniques : Notre cadre répond directement à ce problème en démontrant que des informations cliniquement significatives peuvent être extraites sans données étiquetées, grâce à une combinaison de prompt engineering guidé par schéma et d'une validation hybride. Cette approche est particulièrement avantageuse dans le domaine médical, où l'annotation est coûteuse, chronophage et souvent restreinte pour des raisons de confidentialité.

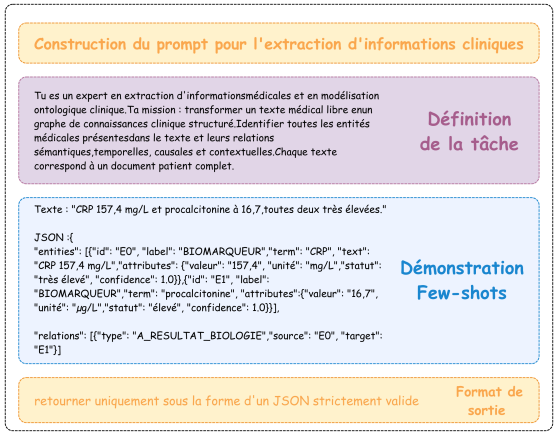

FIGURE 2 – Exemple de prompt. Le prompt comporte trois composantes : **(1) Description de la tâche** encadrée en violet, **(2) Démonstrations few-shot** encadrées en bleu, et **(3) Format de sortie** encadré en jaune.

Nous avons introduit une stratégie d'apprentissage in-context (In Context-Learning, ICL) sans annotation, utilisant des techniques few-shot basées sur des prompts (Figure 2) avec des LLM génériques. Ces modèles peuvent raisonner sur des textes cliniques non annotés et s'adapter dynamiquement à de nouveaux contextes. L'hypothèse centrale est que les LLM peuvent agir comme des annotateurs cliniques autonomes, capturant les relations médicales explicites et implicites sans jeu de données pré-étiqueté, à condition que les prompts soient soigneusement conçus et contraints par des terminologies spécifiques au domaine. La Figure 3 illustre le pipeline d'extraction d'informations

basé sur LLM.

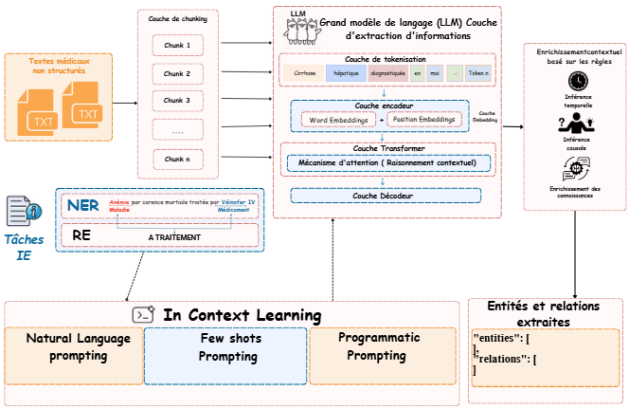

FIGURE 3 – Le pipeline d'extraction d'informations basé sur LLM

In-Context Learning (ICL) : La reconnaissance des entités nommées cliniques et l'extraction de relations ont été réalisées via un LLM utilisant l'ICL. En l'absence de données annotées pour un fine-tuning supervisé, le modèle est guidé par des prompts structurés contenant des exemples annotés manuellement et des instructions explicites pour l'extraction (types d'entités, relations, attributs contextuels). Cette approche permet de générer des sorties structurées spécifiques au domaine sans réentraînement. Grâce à ses mécanismes d'attention, le LLM capture les dépendances sémantiques, infère les relations implicites (par ex. *traité_par*, *causé_par*) et produit des sorties JSON standardisées.

Définition de la tâche : Une stratégie de prompting basée sur les rôles a été adoptée, demandant au LLM d'agir comme un expert en extraction d'informations médicales et en modélisation ontologique clinique. Les catégories d'entités et leur portée sémantique étaient explicitement définies : *MALADIE*, *SYMPTÔME*, *BIOMARQUEUR*, *TRAITEMENT* et *RELATION*. Cette conception exploite les connaissances médicales intégrées du LLM tout en réduisant la complexité du prompt et en améliorant la transférabilité inter-domaines.

Démonstration few-shot : 201 exemples d'entités et de relations ont été utilisés pour le few-shot prompting. Ces exemples constituent des démonstrations contextuelles intégrées dans les prompts, distinctes d'un corpus d'entraînement annoté au sens supervisé du terme : ils ne servent pas à ajuster les paramètres du modèle (fine-tuning), mais à guider son raisonnement par analogie. Chaque exemple comportait une phrase clinique en français et sa représentation structurée JSON avec mentions d'entités, catégories et score de confiance. Nous avons construit ces exemples à partir de phrases extraites des DPI, en annotant les entités et relations selon le schéma ontologique défini. Inspiré du Chain-of-Thought, ce design favorise le raisonnement implicite et améliore la précision sémantique dans les cas complexes (comorbidités, progression de maladies) tout en maintenant les étapes de raisonnement internes.

Format de sortie : Le modèle est contraint de générer des sorties JSON strictement formatées, assurant la validité syntaxique et la compatibilité avec l'intégration dans le graphe de connaissances et les scripts d'évaluation automatisés.

5 Ontologie clinique et raisonnement hybride sur graphe de connaissances

5.1 Conception de l'ontologie et principes de modélisation des connaissances

L'ontologie a été développée selon les principes de développement d'une ontologie 101 [18] pour modéliser les connaissances cliniques liées au diagnostic précoce du sepsis. Construite de zéro mais avec un alignement des termes avec leurs équivalents biomédicaux d'autres ontologies, elle capture symptômes, biomarqueurs, dysfonctionnements d'organes et critères diagnostiques, reliant les observations des patients aux résultats cliniques tout en intégrant le contexte temporel pour soutenir le raisonnement sémantique et l'évaluation du risque de sepsis.

Les ontologies biomédicales existantes n'ont pas été réutilisées en raison de leur complexité, flexibilité limitée et manque d'adaptation aux données cliniques en français. Le modèle définit des contraintes de cardinalité (plusieurs symptômes/biomarqueurs par patient) et des valeurs numériques de biomarqueurs via xsd :float. Des instances issues des DPI ont été validées itérativement à travers des questions de compétence, des vérifications de cohérence logique (HermiT) et des retours d'experts pour garantir la fiabilité clinique et sémantique. La vérification de cohérence logique par HermiT a confirmé l'absence d'inconsistances dans l'ontologie finale.

5.2 Alignement Terminologique

Pour renforcer l'interopérabilité sémantique, les termes harmonisés sont alignés avec des terminologies biomédicales via BioPortal. Le mapping repose sur des embeddings : chaque terme et concept est représenté dans un espace vectoriel commun, et la correspondance est effectuée par similarité cosinus. Nous avons utilisé "paraphrase-multilingual-MiniLM-L12-v2" comme modèle de représentations vectorielles de phrase développé par Sentence-Transformers (HuggingFace). Notre choix se justifie par le fait qu'il soit léger, multilingue et performant pour des tâches de similarité sémantique. Le concept le plus proche est sélectionné, avec un seuil minimal pour garantir la pertinence. Les labels locaux sont conservés pour assurer la traçabilité. Six terminologies, listées au tableau 1 ont été utilisées pour couvrir les domaines cliniques. Ces six terminologies ont été retenues car elles couvrent conjointement les dimensions diagnostiques (SEPON, ICD-9/10, SNOMED CT), pharmacologiques (RXNORM) et de laboratoire (LOINC, MeSH) du sepsis et sont supportées par BioPortal. Cet alignement relie les entités à leurs équivalents biomédicaux canoniques, garantissant qu'elles sont sémantiquement an-

crées dans les terminologies, standardisées pour le raisonnement et la construction du graphe, interopérables avec les ressources cliniques externes, et traçables grâce à un rapport d'harmonisation documentant chaque mapping.

Terminologie	Domaine
SEPON	Concepts cliniques spécifiques au sepsis
ICD-9/10	Maladies et catégories diagnostiques
SNOMED CT	Constatations cliniques et procédures
RXNORM	Médicaments et entités pharmacologiques
LOINC	Tests de laboratoire et diagnostics
MeSH	Termes biomédicaux et de recherche

TABLE 1 – Terminologies utilisées pour l'alignement des entités.

5.3 Intégration dans le Graphe de Connaissances orienté décision

L'ontologie a été transformée en un graphe de connaissances (KG) afin de renforcer la prise de décision clinique, en reliant symptômes, biomarqueurs, événements temporels et résultats diagnostiques dans une représentation cohérente et interprétable. Les entités sont extraites des documents médicaux français via reconnaissance des entités nommées et appariement de motifs, puis normalisées et alignées aux classes ontologiques par similarité lexicale et sémantique. Les relations sont modélisées en triplets RDF (sujet–prédicat–objet), enrichis d'annotations temporelles (relations *precedes*) pour capturer la dynamique des événements, puis sérialisés en RDF/XML et OWL pour le raisonnement et l'interrogation. Ce graphe de connaissances transforme des données cliniques fragmentées en informations exploitables, facilitant le diagnostic précoce du sepsis.

5.4 Architecture du Raisonnement pour l'aide à la décision

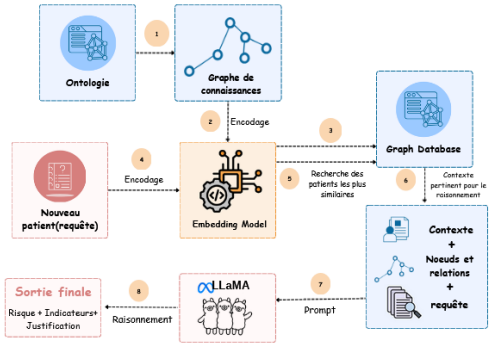

FIGURE 4 – Diagnostic patient utilisant la récupération ontologique et le raisonnement LLM.

Cette étape, illustrée à la figure 4, repose sur un pipeline hybride combinant connaissances structurées issues de l'ontologie et données patient non structurées afin de soutenir des décisions cliniques fiables et interprétables. Les approches purement symboliques sont souvent trop rigides et échouent lorsque les descriptions cliniques s'écartent des règles prédéfinies, tandis que les modèles purement textuels peuvent ignorer des relations médicales essentielles ou produire des inférences non fondées. Afin de pallier ces limitations, notre pipeline unifie raisonnement ontologique, récupération sémantique basée sur des embeddings et raisonnement contextuel guidé par un LLM. L'ontologie constitue le socle du système en structurant les connaissances médicales sous forme de graphe de connaissances. Elle encode les concepts cliniques (symptômes, maladies, signes physiologiques), les relations sémantiques ainsi que des cas patients validés. Le raisonnement s'effectue en parcourant ces relations, permettant d'expliquer explicitement les liens entre facteurs de risque, manifestations cliniques et pathologies. Cette couche symbolique agit comme un ancrage clinique, garantissant la validité médicale des informations manipulées et assurant la traçabilité du processus décisionnel. Afin d'introduire de la flexibilité, les entités de l'ontologie et les descriptions textuelles des patients sont projetées dans un espace vectoriel commun à l'aide d'embeddings hybrides combinant structure du graphe et sémantique linguistique. Cette représentation permet d'effectuer une recherche sémantique avancée et de récupérer des concepts médicaux liés malgré des variations terminologiques, des cas patients historiquement similaires, ainsi que des symptômes et facteurs de risque pertinents dans le contexte du patient. Cette étape joue un rôle de pont entre connaissances structurées et langage clinique réel, compensant le manque de flexibilité du raisonnement symbolique tout en restant ancrée dans l'ontologie.

Le contexte enrichi issu de la phase de récupération est ensuite fourni à un modèle de langage qui produit une interprétation clinique comprenant une estimation du risque de sepsis, l'identification des indicateurs cliniques pertinents et un raisonnement explicatif étape par étape. Afin de garantir la fiabilité des résultats, le modèle est contraint d'utiliser exclusivement les informations récupérées, ce qui limite les hallucinations, assure la cohérence des inférences et maintient l'alignement avec des connaissances médicales validées.

L'intégration de ces composantes permet d'équilibrer rigueur et flexibilité : l'ontologie assure la précision clinique et la structure sémantique explicite, la couche d'embeddings révèle des relations implicites et identifie des profils similaires, tandis que le modèle de langage synthétise ces informations en décisions interprétables. Ce pipeline neuro-symbolique transforme ainsi des données hétérogènes en connaissances actionnables, facilitant un diagnostic précoce du sepsis tout en garantissant un niveau élevé de transparence et de confiance.

6 Expériences et Résultats

6.1 Jeu de Données

Le cadre KG-EDS repose sur un jeu de données privé composé de 58 dossiers patients informatisés (DPI) français réels provenant de patients en soins intensifs dans des hôpitaux parisiens. Ils sont issus de la cohorte APROCCHSS,

comprenant 1 240 patients atteints de sepsis, fournie par l'Assistance Publique–Hôpitaux de Paris (AP-HP), France. Stockés sous forme de PDFs numérisés de dossiers patients, ils intègrent des récits cliniques, des rapports de laboratoire et des résumés d'imagerie. En raison de leur nature hétérogène et non structurée, un pipeline de prétraitement robuste a été développé pour convertir les scans bruts en format lisible par machine et prêt à l'analyse.
Une pseudonymisation des données a été réalisée, avec suppression des identifiants directs (nom, prénom, numéro de sécurité sociale), conformément aux exigences du RGPD et aux protocoles de gouvernance des données médicales de l'AP-HP.
Contrairement aux jeux de données des Unités de Soins Intensifs (USI) typiques limités aux constantes vitales et laboratoires structurés, ce corpus offre une perspective multi-sources textuelles et riche en contexte, capturant l'évolution dynamique du sepsis. Cette diversité permet un raisonnement basé sur ontologie et soutient une détection précoce des risques explicable au-delà des critères diagnostiques conventionnels.

6.2 Résultats pour l'Extraction de Texte Basée sur OCR

Pour évaluer les performances de l'OCR, nous avons extrait une dizaine de rapports manuellement et ils ont été utilisés comme référence. Le tableau 2 résume les résultats sur plusieurs métriques, montrant un haut degré de cohérence entre le texte OCR et les références, avec des taux d'erreur faibles et un alignement sémantique solide.

TABLE 2 – Évaluation des performances OCR selon différentes métriques

Métrique	Score
difflib	0,9342
Levenshtein	0,9668
TF-IDF cosine	0,9784
BLEU	0,8233
ROUGE-L	0,9575
BERTScore	0,9729
WER	0,0915
CER	0,0332

Le système OCR hybride a atteint une qualité de transcription excellente, avec un taux d'erreur de caractères (CER) de 0,0332 et un taux d'erreur de mots (WER) de 0,0915, indiquant que plus de 96% des caractères et mots ont été correctement reconnus. Les métriques de similarité textuelle élevées, telles que Levenshtein (0,9668) et difflib (0,9342), confirment des différences de surface minimales entre la sortie OCR et les références.
Au niveau sémantique, TF-IDF cosine (0,9784) et BERTScore (0,9729) indiquent que le sens et le contexte spécifiques au domaine ont été bien préservés. BLEU (0,8233) et ROUGE-L (0,9575) montrent également un fort alignement des n-grammes et des séquences, avec seulement de légères variations lexicales.

6.3 Évaluation de l'Extraction d'Informations par LLM

6.3.1 Sélection des modèles

Deux LLM légers open-source ont été évalués pour l'extraction d'informations cliniques : LLaMA 3 (8B) et Mistral-Small 3.2, tous deux déployés localement via Ollama pour garantir la confidentialité des données. Ils ont été choisis pour leur efficacité, leur robustesse multilingue et leur fiabilité dans la génération de sorties structurées dans un environnement de déploiement local sans recours à des infrastructures cloud pour des raisons réglementaires (RGPD). LLaMA 3 (8B) dans sa version de base a été privilégié sur la variante instruct afin d'évaluer les capacités de raisonnement du modèle sans ajustement d'instruction préalable, le comportement du modèle de base étant plus contrôlable via le prompt engineering. Des modèles de type transformers supervisés (BioBERT, ClinicalBERT) ont été écartés car ils nécessitent un corpus annoté de fine-tuning spécifique au domaine clinique français, indisponible dans notre contexte
LLaMA 3 (8B) a montré de solides capacités de raisonnement et une conformité constante au schéma JSON requis, facilitant l'intégration dans le graphe de connaissances tout en restant adapté à une exécution locale. Mistral-Small 3.2 a démontré une meilleure gestion du texte médical français, produisant une segmentation plus propre, moins d'erreurs morphologiques et une extraction terminologique plus précise.

6.3.2 Cadre d'évaluation LLM-as-a-Judge

Pour garantir la fiabilité sémantique et clinique des informations extraites automatiquement, nous avons intégré un cadre d'évaluation LLM-as-a-Judge (Figure 5). Dans ce cadre, le même modèle utilisé pour l'extraction est réutilisé comme évaluateur pour juger de la précision, de la cohérence et de l'exhaustivité de ses propres sorties. Le mo-

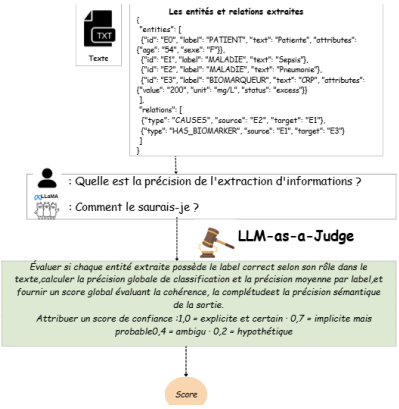

FIGURE 5 – Cadre d'évaluation LLM-as-a-Judge

dèle analyse chaque entité et relation extraites dans leur contexte, détermine si elles remplissent correctement leur rôle sémantique et attribue des scores de confiance reflétant l'explicitation, la certitude et la qualité globale.
Cette approche auto-évaluative, qui présente certaines limites méthodologiques notamment un biais de préférence,

est adoptée en l'absence d'annotations de référence, exploitant les capacités de raisonnement internes du modèle pour évaluer intrinsèquement la précision sémantique et clinique sans besoin de labels manuels.

6.3.3 Résultats pour l'Extraction d'Informations par LLM

TABLE 3 – Comparaison des performances de l'extraction d'information

Métrique	LLaMA	Mistral
Accuracy (Globale)	0,8641	0,9500
Cohérence	0,9024	0,9372
Exhaustivité	0,8481	0,9072
Précision sémantique	0,9410	0,9476

Le tableau 3 présente les résultats consolidés de l'évaluation LLM-as-a-Judge appliquée aux sorties des entités et des relations générées par LLaMA 3 (8B) et Mistral-Small 3.2. Ces métriques évaluent la qualité globale, la cohérence et la fidélité sémantique sur le corpus clinique.
Les deux modèles performent bien, avec des avantages complémentaires. L'accuracy confirme la fiabilité des sorties des entités et des relations, tandis que Mistral se distingue par sa stabilité à travers les dossiers patients. La haute cohérence et exhaustivité montre que les entités et relations restent alignées cliniquement et précises sémantiquement, essentielles pour l'intégration structurée dans le graphe de connaissances.
Globalement, Mistral-Small 3.2 apparaît comme l'option la plus fiable pour le pipeline en aval, grâce à sa meilleure tokenisation du français, son contrôle morphologique et sa clarté sémantique, offrant des frontières d'entités plus nettes et moins d'incohérences, idéal pour la structuration des relations et la construction du KG.

6.4 Évaluation de l'Ontologie

6.4.1 Définition des Questions de Compétence (CQs)

Nous avons formulé les questions de compétence suivantes pour évaluer si l'ontologie capture efficacement les relations entre facteurs cliniques, biologiques et contextuels impliqués dans l'apparition, la progression et la prise en charge du sepsis. Celles-ci ont été validées par deux cliniciens experts dans le domaine du sepsis.

1. CQ1 : Quels facteurs sociaux ou liés au mode de vie sont associés aux patients diagnostiqués avec un sepsis ?
2. CQ2 : Quelles maladies coexistent le plus fréquemment avec le sepsis dans l'ontologie ?
3. CQ3 : Quelles comorbidités augmentent le risque de développer un sepsis ?
4. CQ4 : Quels traitements et médicaments sont le plus souvent administrés aux patients présentant des conditions liées au sepsis ?
5. CQ5 : Quelles chaînes de maladies ou schémas de progression mènent du diagnostic initial au sepsis ou à ses complications ?
6. CQ6 : Quelles conditions médicales ont été identifiées comme causes menant à des effets cliniques liés au sepsis ?

6.4.2 Résultats de l'Ontologie

A l'issue de ce travail, nous avons construit une ontologie KG-EDS comprenant 363 classes et 380 propriétés d'objets, ainsi que divers individus. Son évaluation a été réalisée à l'aide des CQs définies (CQ1–CQ6), visant à tester la couverture, la capacité de raisonnement et l'adéquation à modéliser la nature multifactorielle du sepsis.
Chaque question a été interrogée via des requêtes SPARQL pour déterminer si l'ontologie pouvait extraire ou inférer correctement les informations pertinentes.

6.4.3 Application du raisonnement pour l'aide au diagnostic

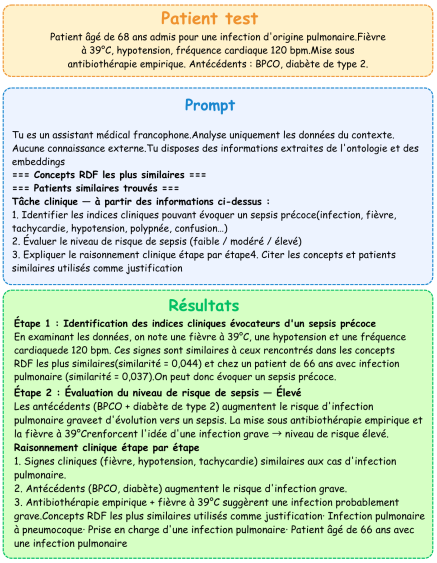

FIGURE 6 – Exemple de diagnostic basé sur LLM

Le pipeline de raisonnement a été testé sur des notes cliniques de patients atteints de sepsis, comme illustré à la figure 6. Le système a associé chaque cas à des concepts ontologiques pertinents, tels que l'infection, la fièvre, l'hypotension et le traitement antibiotique. La recherche fondée sur des embeddings a permis d'identifier des patients et des situations cliniques similaires, contribuant à contextualiser les cas analysés et à mettre en évidence certains schémas potentiels de la maladie. L'étape de raisonnement basée sur un LLM a généré des explications structurées concernant les causes possibles du sepsis, les principaux symptômes observés ainsi qu'une estimation du niveau de risque global (faible, modéré ou élevé). Ces résultats suggèrent que l'approche peut soutenir l'interprétation clinique, tout en nécessitant des validations complémentaires. L'approche semble ainsi offrir un potentiel pour le soutien à la détection précoce, en combinant de manière transparente des connaissances médicales structurées et des informations spécifiques aux patients. Il convient toutefois de préciser que les cas d'usage présentés dans cet article (Figure 6) cor-

respondent à des scénarios simulés et n'ont pas encore été évalués sur des patients réels.

7 Conclusion et Perspectives Futures

Nous avons proposé **KG-EDS (Knowledge Graph for Early Diagnosis of Sepsis)** qui repose sur trois étapes principales. D'abord, une préparation robuste des données traite l'hétérogénéité des dossiers cliniques (manuscrits, numérisés ou mal formatés) grâce à un pipeline OCR hybride combinant Donut Transformer et Tesseract, assurant une extraction fiable malgré la qualité variable des documents. Ensuite, les informations cliniques sont extraites par LLM via une stratégie d'in-context learning et de prompt engineering, permettant d'identifier symptômes, biomarqueurs et traitements sans grands jeux de données annotés. Pour limiter hallucinations et biais, les résultats sont alignés sur des terminologies biomédicales et validés par des règles. Enfin, les entités sont intégrées dans un graphe de connaissances basé sur des terminologies standardisées (SNOMED CT, ICD-10, UMLS), offrant une représentation interprétable, interopérable et exploitable pour le raisonnement clinique structuré.

Il est important de noter que, bien que testé dans le contexte du sepsis sur un dataset réduit, cette méthodologie est intrinsèquement *généralisable*. Le pipeline OCR, la stratégie d'extraction basée sur LLM, le processus d'alignement ontologique et la construction du graphe de connaissances sont des composants indépendants de la pathologie. En remplaçant l'ontologie du domaine et en adaptant les prompts d'extraction à une autre pathologie, le même pipeline pourrait, par exemple, être appliqué aux maladies cardiovasculaires ou aux maladies infectieuses.

Malgré des résultats prometteurs, certaines limites subsistent. Les performances du système dépendent de la qualité de la documentation, et les hallucinations, bien que réduites, ne sont pas complètement éliminées. Les erreurs de classification et le chevauchement des catégories d'entités indiquent la nécessité d'un meilleur modelage contextuel et d'un ajustement ciblé. De plus, le potentiel biais des LLM et la représentativité limitée du dataset imposent une vigilance particulière lors de l'interprétation des résultats et de la généralisation à d'autres populations cliniques.

Les travaux futurs viseront à renforcer la robustesse et la fiabilité de KG-EDS selon plusieurs axes complémentaires : l'enrichissement et la diversification des datasets avec des annotations validées par des cliniciens pour réduire les biais et établir un benchmark de référence ; la mise en place de stratégies avancées, notamment des prompts en ensemble et des mécanismes d'atténuation, afin de renforcer la cohérence factuelle et limiter les hallucinations des LLM ; le développement du raisonnement temporel inter-documents pour mieux suivre la progression des maladies ; et l'optimisation des algorithmes de raisonnement sur graphe pour assurer la scalabilité clinique et le déploiement en temps réel, tout en évaluant systématiquement la généralisation du pipeline à d'autres pathologies.

Remerciements

Ce travail est financé et soutenu par : le programme Horizon Europe de l'Union européenne (MSCA, convention n° 101236749), les financements France 2030 RHU RECORDS (ANR-18-RHUS-0004), IHU SEPSIS (ANR-23-IAHU-0004), iRECORDS financé par ERA PerMed (JTC_2021), le Programme d'Investissements d'Avenir (I-SITE ULNE / ANR-16-IDEX-0004 ULNE), géré par l'Agence Nationale de la Recherche (n° I-KUL-22-005-ARCHIE-INFINITE), ainsi que d'un financement de l'Inserm et du Ministère de la Santé via l'appel MESSIDORE 2023 opéré par l'IReSP (AAP-2023-MSDR-341423). Il a été réalisé dans le cadre d'un stage de Master 2 Recherche de 6 mois et financé par l'institut DATAIA Paris Saclay.

Références

[1] Fatemeh Amrollahi, Supreeth P Shashikumar, Fereshteh Razmi, and Shamim Nemati. Contextual embeddings from clinical notes improves prediction of sepsis. In *AMIA annual symposium proceedings*, volume 2020, page 197, 2021.

[2] Mercedes Arguello Casteleiro, Diego Maseda Fernandez, George Demetriou, Warren Read, Maria Jesus Fernandez Prieto, Julio Des Diz, Goran Nenadic, John Keane, and Robert Stevens. A case study on sepsis using PubMed and deep learning for ontology learning. *Studies in Health Technology and Informatics*, 235 :516–520, 2017.

[3] Sabrina Azzi. *Nouvelle méthodologie de construction d'ontologies médicales : cas d'étude : diagnostic de la pneumonie*. PhD thesis, Université du Québec en Outaouais, Gatineau, Canada, 2021.

[4] BioPortal. Sepsis ontology (SEPON). https://bioportal.bioontology.org/ontologies/SEPON, 2024.

[5] Felix Brann, Nicholas William Sterling, Stephanie O Frisch, and Justin D Schrager. Sepsis prediction at emergency department triage using natural language processing : Retrospective cohort study. *JMIR AI*, 3 :e49784, 2024.

[6] Elena Calvo-Cidoncha, Concepción Camacho-Hernando, Faust Feu, Xavier Pastor-Duran, Carles Codina-Jané, and Raimundo Lozano-Rubí. OntoPharma : Ontology based clinical decision support system to reduce medication prescribing errors. *BMC Medical Informatics and Decision Making*, 22(1) :238, 2022.

[7] Laura Forero Camacho, Nacéra Seghouani, Gregoire Tournois, and Farida Zehraoui. OntoSepsisKG : Omics-based knowledge graph for sepsis. In *Proceedings of the Conference on Knowledge Graphs and Ontologies in Biomedicine*, 2024.

[8] Kevin Donnelly. Snomed-ct : The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121 :279–290, 2006.

[9] Sudhakaran Gajendran, D Manjula, Vijayan Sugumaran, and R Hema. Extraction of knowledge graph of Covid-19 through mining of unstructured biomedical corpora. *Computational Biology and Chemistry*, 102 :107808, 2023.

[10] Authia P Gray, Erin Chung, Nasir Abbas, Ibrahim Elsohaby, Ginenus Fekadu, Mohsen Naghavi, GBD 2021 Global Sepsis Collaborators, et al. Global, regional, and national sepsis incidence and mortality, 1990–2021 : a systematic analysis. *The Lancet Global Health*, 13(12) :e2013–e2026, 2025.

[11] Ayoub Harnoune, Maryem Rhanoui, Mounia Mikram, Siham Yousfi, Zineb Elkaimbillah, and Bouchra El Asri. BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Computer Methods and Programs in Biomedicine Update*, 1 :100042, 2021.

[12] Peter J Haug, Jeffrey P Ferraro, John Holmen, Xinzi Wu, Kumar Mynam, Matthew Ebert, Nathan Dean, and Jason Jones. An ontology-driven, diagnostic modeling system. In *Journal of the American Medical Informatics Association*, volume 20, pages e102–e110, 2013.

[13] Yongqun He, Hong Yu, Edison Ong, Yang Wang, Yingtong Liu, Anthony Huffman, Hsin-hui Huang, John Beverley, Junguk Hur, Xiaolin Yang, et al. Cido, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Scientific data*, 7(1) :181, 2020.

[14] Khandaker Reajul Islam, Johayra Prithula, Jaya Kumar, Toh Leong Tan, Mamun Bin Ibne Reaz, Md Shaheenur Islam Sumon, and Muhammad EH Chowdhury. Machine learning-based early prediction of sepsis using electronic health records : A systematic review. *Journal of Clinical Medicine*, 12(17) :5658, 2023.

[15] Xia Jing, Hua Min, Yang Gong, Paul Biondich, David Robinson, Timothy Law, Christian Nohr, Arild Faxvaag, Lior Rennert, Nina Hubig, et al. Ontologies applied in clinical decision support system rules : Systematic review. *JMIR Medical Informatics*, 11, 2023.

[16] Natalja Kurbatova and Rowan Swiers. Disease ontologies for knowledge graphs. *BMC Bioinformatics*, 22(1) :377, 2021.

[17] Lucas Gomes Maddalena, Fernanda Araujo Baião, Tiago Prince Sales, and Giancarlo Guizzardi. Evaluating ontologically-aware large language models : An experiment in sepsis prediction. *Semantic Web*, 0(0) :1–23, 2025.

[18] Natalya F Noy and Deborah L McGuinness. Développement d'une ontologie 101 : Guide pour la création de votre première ontologie. *Université de Stanford, Stanford, Traduit de l'anglais par Anila Angjeli. ht tp ://www. bnf. fr/pages/infopro/normes/pdf/noDevOnto. pdf*, 2000.

[19] C. Piriou, S. Despres, J. Nobecourt, C. Irles, and C. Le Roy. Graphe de connaissance et ontologie pour la représentation des données de la LLC. In *Journée Santé et IA 2023, PFIA 2023 (AFIA)*, 2023.

[20] Xia Qiu, Yu-Peng Lei, and Rui-Xi Zhou. SIRS, SOFA, qSOFA, and NEWS in the diagnosis of sepsis and prediction of adverse outcomes : A systematic review and meta-analysis. *Expert Review of Anti-Infective Therapy*, 21(8) :891–900, 2023.

[21] Venkat Reddy, Harshitha Reddy, Rinkle Gemnani, Sunil Kumar, and Sourya Acharya. Navigating the complexity of scoring systems in sepsis management : A comprehensive review. *Cureus*, 16(2) :e54030, 2024.

[22] Regenstrief Institute. LOINC : Logical observation identifiers names and codes. https://loinc.org/, 2025. Version 2.81, consulté le 1er mai 2026.

[23] Adam R Schertz, Kristin M Lenoir, Alain G Bertoni, Beverly J Levine, Morgana Mongraw-Chaffin, and Karl W Thomas. Sepsis prediction model for determining sepsis vs SIRS, qSOFA, and SOFA. *JAMA Network Open*, 6(8) :e2329729, 2023.

[24] Manu Shankar-Hari, Gary S Phillips, Mitchell L Levy, Christopher W Seymour, Vincent X Liu, Clifford S Deutschman, Derek C Angus, Gordon D Rubenfeld, and Mervyn Singer. Developing a new definition and assessing new clinical criteria for septic shock. *JAMA*, 315(8) :775–787, 2016.

[25] Mervyn Singer, Derek C Angus, Djillali Annane, Michael Bauer, Andre C Kalil, Michael Klompas, Flavia R Machado, Greg S Martin, Adrienne G Randolph, Manu Shankar-Hari, et al. Sepsis. *The Lancet*, 407(10535) :1276–1288, 2026.

[26] U.S. National Library of Medicine. Medical subject headings (MeSH). https://www.nlm.nih.gov/mesh/meshhome.html, 2025. Consulté le 1er mai 2026.

[27] World Health Organization. *International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10)*, 2016.

[28] Melissa Y Yan, Lise Tuset Gustad, Lise Husby Høvik, and Øystein Nytrø. Terminology and ontology development for semantic annotation : A use case on sepsis and adverse events. *Semantic Web*, 14(4) :681–744, 2023.

[29] Melissa Y Yan, Lise Tuset Gustad, and Øystein Nytrø. Sepsis prediction, early detection, and identification using clinical text for machine learning : A systematic review. *Journal of the American Medical Informatics Association*, pages 1–17, 2021.

[30] Hao Yang, Jiaxi Li, Chi Zhang, Alejandro Pazos Sierra, and Bairong Shen. Large language model–driven knowledge graph construction in sepsis care using multicenter clinical databases : Development and usability study. *Journal of Medical Internet Research*, 27 :e65537, 2025.