Récentes avancées de l'inférence en langue naturelle pour les essais cliniques

M. Aguiar[1], P. Zweigenbaum[1], N. Naderi[1]

[1] Université-Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique

{mathilde.aguiar, pierre.zweigenbaum, nona.naderi}@lisn.upsaclay.fr

Résumé

Cet article présente une revue de la littérature pour la tâche d'inférence en langue naturelle (Natural Language Inference en anglais) appliquée aux essais cliniques. La tâche d'inférence en langue naturelle implique différents niveaux de raisonnement, à la fois sur le plan sémantique, numérique et en utilisant des connaissances du monde réel. Son application aux données de santé, en particulier aux essais cliniques, amène un défi supplémentaire lié au vocabulaire spécifique. Dans nos travaux nous nous concentrons sur les récentes avancées dans ce domaine en réalisant une revue de la littérature parue entre 2022 et 2023. En appliquant la méthode PRISMA, nous avons analysé 74 études issues de diverses bases de données.

Mots-clés

Inférence en langue naturelle, Essai clinique, Traitement automatique des langues, Revue de la littérature.

Abstract

This paper presents a review of the recent literature on Natural Language Inference (NLI) applied to clinical trials. NLI implies different levels of reasoning, both on a semantic level, numerical reasoning or using real-world knowledge. Applying it on medical data, especially on clinical trials, involves further challenges due to domain-specific vocabulary. In our work, we focus on recent changes in the domain by conducting a literature review of studies published between 2022 and 2023. Using the PRISMA framework, we analyzed 74 studies taken from different databases.

Keywords

Natural Language Inference, Clinical Trial, Natural Language Processing, Literature review.

1 Introduction

La tâche de Reconnaissance d'Implication Textuelle (*Recognising Textual Entailment* (RTE) en anglais) a été popularisée lors de la campagne d'évaluation PASCAL [12] . La tâche d'inférence en langue naturelle (ILN) (*Natural Language Inference* (NLI) en anglais) est équivalente à celle de RTE et a notamment été formalisée par [32]. Le but de l'ILN est de déterminer si pour une prémisse donnée, l'hypothèse considérée peut être inférée depuis cette prémisse (*entailment* en anglais) ou s'il existe une contradiction entre celles-ci.

Hypothèse : « *Le patient ne souffre pas de diabète.* »,
Prémisse : « *Le patient s'est injecté 1 unité d'insuline à 7 heures.* » → **Contradiction**.

Nous étudions ici l'application de l'ILN au domaine des essais cliniques. Un essai clinique a pour but d'étudier l'efficacité d'un nouveau traitement ou protocole développé précédemment en laboratoire. Pour ce faire, les chercheurs ont besoin de recruter une population de patients ayant un profil clinique bien défini. Les essais cliniques comprennent de nombreuses informations telles que : les critères d'éligibilité, les interventions médicales effectuées, les effets secondaires ainsi que les résultats. L'ILN peut être appliquée aux essais cliniques en considérant le rapport d'essai clinique en tant que prémisse. L'hypothèse peut quand à elle illustrer différentes situations telles que le profil d'un patient ou encore une affirmation portant sur les résultats obtenus lors de l'essai clinique.

Cette revue a pour but de synthétiser les différentes études sur le sujet parues entre 2022 et 2023 pour établir un état de l'art des techniques employées et des ressources disponibles tout en déterminant comment l'ILN pourrait être bénéfique aux essais cliniques, ainsi que déterminer les défis et perspectives du domaine. Dans la suite de cet article, nous présentons des travaux et revues antérieures sur l'ILN et le TAL dans le domaine clinique (sec. 2), les méthodes employées pour réaliser notre revue de la littérature (sec. 3), et la synthèse des articles étudiés, en commençant par le domaine général, pour se spécialiser vers l'ILN dans le domaine clinique (sec. 4). Nous résumons enfin nos contributions et des pistes de travaux futurs (sec. 5).

2 Travaux connexes

D'autres revues de la littérature se sont intéressées aux applications du traitement automatique des langues (TAL) dans un contexte clinique, ainsi qu'aux avantages que le TAL pourrait apporter au domaine. Gao *et al.* [15] regroupent les différentes tâches de TAL disponibles dans le domaine clinique en analysant 35 études de 2007 à 2021 recensées en considérant les bases de données suivantes : PubMed, Embase, ACM Digital Library, WebOfScience et ACL Anthology. Crema *et al.* [11] se concentrent sur les ressources et techniques pour l'extraction d'information,

les tâches de classification et l'inférence de données pour le traitement des dossiers médicaux informatisés dans le domaine des neurosciences et de la psychiatrie. Idnay *et al.* [24] ont fait une revue de onze études présentant des systèmes d'aide au recrutement de patients dans des essais cliniques. Dans le domaine général, Putra *et al.* [38] a fait une large revue des applications, jeux de données, approches et défis pour le RTE, en analysant 274 études parues entre 2012 et 2021. Notre contribution se distingue des précédentes en examinant des études plus récentes publiées en 2022 et 2023 et en se focalisant sur la tâche d'ILN.

3 Méthodes

Pour sélectionner les articles de cette revue nous avons suivi la méthode de référence *Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA)* [34] (voir fig. 1). Nous avons sélectionné des bases de données en accord avec notre domaine : PubMed, ACL Anthology, Science Direct, Scopus, ACM, DBLP, Web Of Science et IEEE. Nous avons ensuite défini les requêtes suivantes que nous avons exécutées sur les moteurs de recherche de ces bases de données :

1. clinical AND "Natural Language Processing" AND "Natural Language Inference" OR NLI
2. clinical AND "Textual Entailment" OR TE
3. "Natural Language Inference"

Textual Entailment étant un autre terme pour l'ILN, nous incluons une requête avec ce mot clé. Nous récoltons un total de 1921 articles pour ces trois requêtes. Après avoir supprimé les 207 doublons, nous obtenons 1714 articles. Pour préciser notre requête nous définissons les critères d'inclusion et exclusion suivants, que nous vérifions manuellement :

— Inclusion :
  — Apprentissage profond (*Deep Learning*)
  — Apprentissage automatique (*Machine Learning*)
  — Évaluation par des pairs
  — Publié en anglais
  — Publié entre 2022 et 2023
— Exclusion :
  — Multi-modalité, traitement de l'image
  — Étude multi-tâches/non focalisée sur l'ILN

1640 articles sont exclus après avoir lu leurs titres et résumés et en ne considérant que les études parues entre 2022 et 2023, laissant 74 articles. Un article a été rejeté après lecture complète. Un article a été ajouté via la relève de citations dans les articles analysés. Après lecture complète, 74 articles sont retenus pour notre revue de la littérature. Parmi ces 74 articles, 23 portent sur la campagne d'évaluation NLI4CT qui traite des essais cliniques en particulier. Huit études portant sur le domaine clinique mais ne traitant pas d'essais cliniques sont également inclues dans notre revue. Compte tenu du leur faible nombre, nous considérons que cette étude porte principalement sur les essais cliniques. Un tableau récapitulatif des études incluses dans cette revue est disponible sur notre GitHub [1] et en annexe dans les tableaux 1, 2 et 3 [2].

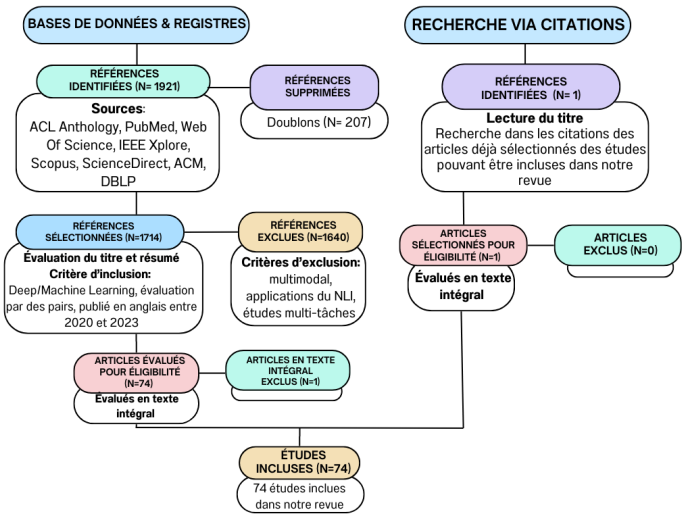

FIGURE 1 – Diagramme de flux PRISMA

Nous cherchons à répondre aux questions de recherche suivantes :

— Quels sont les méthodes, jeux de données et modèles employés pour résoudre la tâche d'inférence en langue naturelle pour le domaine clinique (sec. 4.1 et 4.2) ?
— Comment l'inférence en langue naturelle pourrait-elle bénéficier au domaine des essais cliniques (sec. 4.3) ?
— Quels sont les défis à relever pour l'inférence en langue naturelle appliquée aux essais cliniques (sec. 4.4) ?

4 Résultats

Nous rapportons ici les réponses obtenues aux questions posées ci-dessus.

4.1 Inférence en langue naturelle dans le domaine général

Définition L'inférence en langue naturelle (ILN) fait partie des tâches d'une famille souvent désignée comme « Compréhension de la langue naturelle » (*Natural Language Understanding* (NLU) en anglais). Le modèle doit apprendre une fonction

$$f(h_i, p_i) \rightarrow y_i$$

où h_i est une hypothèse, p_i une prémisse et $y \in e, c, n$ est la relation à prédire parmi e (*entailment*), c (*contradiction*) et n (*neutral*).

1. https://github.com/MathildeAguiar/CTInfer/blob/main/recentes_avancees_NLI_essais_cliniques/etudes_2022-2023.md

2. Par ailleurs, dans cet article, nous distinguons ces travaux en les citant selon le format [auteur, année] qui fait référence à une bibliographie séparée nommée *Études incluses*.

Pour déterminer la relation à prédire, le modèle doit souvent faire appel à plusieurs types d'inférences et connaissances : (a) inférence numérique, (b) connaissances du monde réel et « bon sens ». (a) implique que le modèle est capable de raisonner sur des quantités (savoir faire des conversions/équivalences etc.), reconnaître des valeurs numériques et réaliser tout type de calcul. (b) concerne toutes les notions de « bon sens » que le modèle doit mobiliser, relatives à la connaissance du monde réel, etc.

Tâches et jeux de données existants Pour évaluer les modèles développés, certains travaux mettent à disposition des jeux de données, et définissent des « tâches » à réaliser. Ces jeux de données peuvent également être mis à disposition lors de campagnes d'évaluation. Les tâches peuvent traiter d'un aspect linguistique en particulier (comme la négation [18, 22], la présupposition et l'implicature [26], etc.) ou encore être destinées à évaluer les capacités des modèles sur un type de document, etc. Dans le contexte de notre revue nous nous limitons aux jeux de données publiés entre 2022 et 2023. Dans le domaine général, nous observons un effort apporté pour publier des jeux de données dans d'autres langues que l'anglais : arabe [Al Jallad et Ghneim, 2023], vietnamien [Huynh *et al.*, 2022], certaines langues indiennes [Aggarwal *et al.*, 2022] et même pour des langues très peu dotées en ressources linguistiques informatisées comme le créole jamaïcain [Armstrong *et al.*, 2022] et les langues indigènes d'Amérique [Kann *et al.*, 2022]. Certains de ces jeux de données sont créés à partir de données originales en langue cible (par exemple en extrayant les prémisses d'articles de journaux, sites gouvernementaux, etc.) [Huynh *et al.*, 2022]. D'autres approches [Kumar Upadhyay et Kumar Upadhya, 2023, Aggarwal *et al.*, 2022] utilisent des outils de traduction automatique, associés à une vérification manuelle, pour traduire des jeux de données pré-existants comme XNLI [10]. Cette dernière approche permet d'obtenir des jeux de données plus conséquents mais souvent au détriment de la qualité des exemples produits.

D'autres tâches s'intéressent à un phénomène linguistique en particulier, comme par exemple [Truong *et al.*, 2022] qui a développé un jeu de données pour évaluer la capacité des modèles de langue à raisonner en présence de négations. [Liu *et al.*, 2022] propose une nouvelle façon de construire des jeux de données pour l'ILN en générant des prémisses supplémentaires à partir d'un jeu de données existant en utilisant GPT-3 [6]. La majorité des jeux cités précédemment sont constitués d'une prémisse, pouvant aller d'une simple phrase à un document complet, d'une hypothèse, généralement une seule phrase, et du label associé.

Par ailleurs, il existe quelques jeux de données propres à certains domaines : domaine clinique (voir Sec 4.2), scientifique en général [Sadat et Caragea, 2022] ou juridique [27].

Différentes approches Traditionnellement, l'une des approches la plus courante pour l'ILN est basée sur des modèles discriminants (modèles de langue masqués, MLM) comme BERT [13], RoBERTa [31], DeBERTa [19] ou encore des Sentence-Transformers comme Sentence-BERT [41]. Ces modèles sont utilisés comme encodeurs devant un classifieur que l'on entraîne sur une tâche donnée, en affinant éventuellement le modèle en même temps (*finetuning*) [Gubelmann *et al.*, 2023, Liu *et al.*, 2022]. Bien que moins courant, il est aussi possible d'affiner des modèles génératifs pré-entraînés et de leur apposer une couche finale linéaire pour réaliser les prédictions sur la tâche d'ILN [Huang *et al.*, 2023].

D'autres approches utilisent l'apprentissage contrastif en entraînant leur modèles sur des paires contenant des exemples positifs et négatifs et en essayant de minimiser à la fois la fonction de perte de ces exemples contrastifs et celle pour la classification [Mersinias et Valvis, 2022, Li *et al.*, 2022, Li *et al.*, 2023, Corrêa Dias *et al.*, 2023, Feng *et al.*, 2023].

Pour augmenter la performance des résultats obtenus suite à un entraînement, [Vladika et Matthes, 2023, Zhou *et al.*, 2023] utilisent un ensemble de plusieurs modèles (le même modèle entraîné avec des amorces aléatoires différentes ou des modèles totalement différents) qu'ils font voter à la majorité pour déterminer le label à prédire. Lorsque la prémisse est un texte de plusieurs phrases, déterminer quelles phrases contiennent des indices permettant de déterminer la relation peut être utile. [Vladika et Matthes, 2023] emploie pour cela un apprentissage multi-tâche qui apprend simultanément à déterminer quelles phrases contiennent des indices et quelle relation existe entre prémisse et hypothèse. D'autres systèmes insèrent cette étape de classification des phrases avant [Alameldin et Williamson, 2023, Bevan *et al.*, 2023] ou après [Zhou *et al.*, 2023, Zhao *et al.*, 2023] la tâche d'ILN proprement dite, en utilisant deux modèles distincts.

Comme mentionné plus haut, les jeux de données peuvent être de taille restreinte, en particulier pour une langue à faibles ressources ou un domaine spécifique. [Kann *et al.*, 2022] tirent avantage de l'apprentissage translingue ou multilingue qui permet aux langues peu dotées de bénéficier de connaissances apprises par un modèle entraîné sur des langues bénéficiant de beaucoup de données.

Méthodes d'évaluation La majorité des systèmes sont évalués en utilisant la mesure F1, la précision, le rappel ou l'exactitude (*accuracy*). Certaines études cherchent également à dépister un problème présent dans certains jeux de données où la formulation de l'hypothèse permet souvent à elle seule de prédire la relation [37]. La performance d'un modèle entraîné uniquement sur l'hypothèse donne alors une base de comparaison qui permet de quantifier ce biais [Asael *et al.*, 2022, Armstrong *et al.*, 2022, Saxon *et al.*, 2023, Jullien *et al.*, 2023a].

Domaines d'application La tâche d'ILN peut servir de base pour la réalisation de tâches « de plus haut niveau » comme le résumé automatique [28, 5] ou la recherche de réponses à des questions [2]. Dans ces cas, l'ILN peut être utilisée lors d'un apprentissage par transfert, soit en utilisant directement un modèle déjà pré-entraîné sur une tâche d'ILN [Kann *et al.*, 2022], [36], soit en prolongeant son ap-

prentissage sur la tâche visée. Une autre approche est d'utiliser le jeu de données d'ILN en combinaison avec une ou plusieurs tâches cibles lors d'un apprentissage multi-tâche [8].

4.2 Méthodes, modèles et jeux de données dans le domaine clinique

Jeux de données Le jeu de données BioNLI [Bastan *et al.*, 2022] est construit automatiquement à partir de résumés de publications PubMed [3], où la prémisse décrit une expérience scientifique et contient les éléments pour en déduire la relation d'inférence, alors que l'hypothèse résume l'expérience décrite dans la prémisse. Les exemples originaux extraits depuis PubMed sont tous des exemples positifs (labellisés *entailment*). Les exemples négatifs sont construits automatiquement en transformant des exemples positifs en utilisant des règles comme l'« inversion d'entités nommées ». [Bastan *et al.*, 2022] ont entraîné un classifieur en utilisant uniquement l'hypothèse en entrée et obtiennent un score F1 moyen de 0.63, soit seulement 10 points de moins que lorsque l'hypothèse et la prémisse sont utilisées pour l'entraînement, ce qui indique la présence d'artefacts.

NLI4CT regroupe une collection d'essais cliniques sur des traitements pour le cancer du sein issus de clinicaltrials.gov. Ceux-ci vont constituer les prémisses du jeu de données. Ils possèdent les sections suivantes : *Éligibilité*, *Intervention*, *Effet secondaires* et *Résultats*. Il existe deux types d'exemples : *Simple*, et *Comparaison*, où deux documents sont mis en parallèle pour pouvoir déduire leur relation avec une hypothèse donnée. Les hypothèses sont créées par des experts cliniciens et un soin particulier a été apporté au fait de ne pas créer d'hypothèses triviales, en diversifiant les styles et les tournures de phrases employées. Le but est d'évaluer différents types d'inférence : connaissance du monde réel, numérique et biomédicale (comme les synonymes, les relations taxonomiques et les acronymes). Malgré le soin apporté à la création des hypothèses et à l'annotation, NLI4CT souffre tout de même de la présence d'artefacts.

Méthodes La campagne d'évaluation NLI4CT [Jullien *et al.*, 2023b] de SemEval 2023 a incité le développement de nombreux systèmes et a ainsi diversifié les approches cherchant à traiter cette tâche. Une partie des systèmes [Vassileva *et al.*, 2023, Volosincu *et al.*, 2023, Wang *et al.*, 2023] tirent parti d'un apprentissage par transfert en utilisant des modèles MLM pré-entraînés sur des tâches ou données biomédicales ou cliniques comme ClinicalBERT [23], BioBERT [29] et bien d'autres. De nouvelles approches utilisent des grands modèles de langue autorégressifs (*Large Language Models*, LLM) [Zhao *et al.*, 2023, Pahwa et Pahwa, 2023, Rajamanickam et Rajaraman, 2023,
Kanakarajan et Sankarasubbu, 2023]. Ils les amorcent par un texte (*prompt*) qui contient généralement des instructions et éventuellement quelques exemples et solutions associées (démonstrations) (*Few-Shot Learning*) ou pas (*Zero-Shot Learning*). [Zhao *et al.*, 2023] utilisent le concept de chaîne de pensée [45] dans ses démonstrations pour décortiquer le raisonnement à suivre pour résoudre la tâche. [Rajamanickam et Rajaraman, 2023] utilisent des règles en post-génération pour étayer la réponse produite par le modèle. [Kanakarajan et Sankarasubbu, 2023] affinent Flan-T5 [9] en utilisant leurs propres instructions (*instruction-tuning*).

Pour pallier le manque de données, [Feng *et al.*, 2023, Takehana *et al.*, 2023] proposent plusieurs techniques d'augmentation de données, comme perturber de façon aléatoire les données, créer de nouveaux exemples en remplaçant certains termes par leurs synonymes ou encore traduire les exemples dans une langue étrangère puis les re-traduire dans leur langue d'origine. Cette technique a permis à [Feng *et al.*, 2023] de gagner quelques points de score F1, mais comparé aux autres systèmes soumis pour la campagne d'évaluation, ces deux systèmes ont une moins bonne performance que les autres. [Alameldin et Williamson, 2023] poursuivent le pré-entraînement de leur modèle initial sur des documents supplémentaires issus de clinicaltrials.gov pour obtenir des modèles entraînés sur des données similaires à la tâche cible.

[Conceição *et al.*, 2023] utilisent plusieurs ontologies médicales comme l'Ontology of Adverse Events (OAE) [20] pour annoter des entités. En plus des annotations, ils utilisent des règles de simplification et de conversion numérique. La prédiction est ensuite fondée sur un score de similarité entre prémisse et hypothèse. [Noor Mohamed et Srinivasan, 2023] utilisent aussi une combinaison d'informations issues de bases de connaissances et d'un ensemble de règles sémantiques : règle pour la double négation, pour la déduction, pour les conditions, etc. L'ajout d'annotations supplémentaires et de règles ont augmenté la performance des modèles mais en comparaison à d'autres approches évaluées sur NLI4CT, ces approches restent parmi les moins performantes.

Modèles Certaines approches utilisent des MLM pré-entraînés sur des données du domaine général comme BERT, DistilBERT [43], RoBERTa, StructBERT [44], DeBERTa, SBERT ou Longformer [4]. Dans l'édition 2023 de NLI4CT, de nombreuses équipes ont employé des LLM, aussi bien des décodeurs comme GPT2 [39], GPT-3.5, ChatGPT que des modèles séquence-à-séquence comme BART [30], T5 [40] ou sa version affinée sur des instructions Flan-T5 [9], ou encore sa version affinée sur des articles scientifiques SciFive [35].

D'autres approches préfèrent employer des modèles pré-entraînés sur des données biomédicales comme BioBERT, BioLinkBERT [46] ou SciBERT [3]. Leur entraînement étant sur un vocabulaire proche de celui du vocabulaire clinique, cela peut apporter un gain de performance. Enfin, UmlsBERT [33], ClinicalBERT et BioClinicalBERT [1] ont été pré-entraînés sur des dossiers médicaux électroniques et PubMedBERT [16] a été pré-entraîné sur des

3. https://pubmed.ncbi.nlm.nih.gov/

publications issues de PubMed.

4.3 Bénéfices de l'ILN pour le domaine des essais cliniques

Les essais cliniques se déroulent généralement sur plusieurs années et sont un processus coûteux. La numérisation de ces essais cliniques est réalisée dans le but d'automatiser certaines phases du processus et de faciliter l'accès aux résultats et autres données dont les chercheurs pourraient avoir besoin. Inan *et al.* [25] relèvent trois niveaux auxquels cette numérisation cherche à contribuer : le recrutement de patients, la collecte de données de santé et l'analyse des données collectées. L'objectif visé est notamment de réduire les coûts liés au recrutement des patients ; de recruter des patients ayant des profils plus variés tout en correspondant aux critères d'éligibilité définis pour l'essai clinique ; de simplifier le traitement et l'exploitation des résultats obtenus [42].

Zhang *et al.* [48] proposent une méthode pour le recrutement de patients en modélisant leur dossier médical électronique et les critères d'éligibilité d'un essai clinique donné en les traduisant en une tâche d'ILN. Les documents patients et les critères d'éligibilité sont modélisés en utilisant des vecteurs de mots distincts basés sur BERT et ClinicalBERT. Les auteurs appliquent ensuite un module de raisonnement numérique puis détermine la relation entre prémisse et hypothèse. Dhayne et Kalani [14] proposent une solution similaire mais plutôt que de prédire une relation (*entailment* ou *contradiction*), ils calculent un score de similarité continu entre les données d'un patient et les critères d'éligibilité. [Widiana *et al.*, 2022] et [Sosa *et al.*, 2023] proposent tous deux d'utiliser l'ILN pour exploiter la littérature liée au COVID-19. [Widiana *et al.*, 2022] ont développé un système permettant de valider une affirmation donnée en la confrontant à la littérature existante. Pour ce faire ils détectent les informations dans les études considérées qui supportent cette affirmation puis, en fonction de cela, prédisent s'il y a une implication. [Sosa *et al.*, 2023] utilisent des paires de phrases issues de la littérature et cherchent à déterminer si les phrases d'une même paire se contredisent. Le but étant qu'en période de pandémie, pour trouver un traitement adéquat, les chercheurs devaient se fonder sur des études dont la qualité pouvait être variable ; ce système de détection de contradiction visait à aider à étayer le choix du traitement à administrer. [Chen *et al.*, 2023] ont développé un système de détection de symptômes et du statut associé à ce symptôme (si le patient souffre ou non de celui-ci). Le modèle utilise à la fois l'hypothèse, la prémisse et la définition du symptôme concerné pour déterminer quel label prédire. Ce type de système pourrait être utile lorsque l'on cherche à déterminer si un patient convient aux critères d'éligibilité ou encore pour détecter des effets secondaires suite à la prise d'un traitement.

4.4 Défis et perspectives

La présence d'artefacts dans les jeux de données est un problème récurrent à la fois dans le domaine général et dans les jeux de données cliniques [21]. Ces artefacts vont donner des indications au modèle pour prédire un label sans même tenir compte de la relation réelle entre l'hypothèse et/ou la prémisse. Certains mots sont en effet plus souvent utilisés dans les hypothèses correspondant à un label en particulier [17], par exemple l'emploi d'adverbes de négation pour les instances labélisées en tant que *contradiction* ou encore l'utilisation de mot génériques tels que *animal* ou *instrument* pour les instances labélisées *entailment*. La longueur de l'hypothèse est aussi une forme d'artefact, avec les hypothèses des instances labélisées *entailment* généralement les plus courtes et celles *neutre* les plus longues. Pour détecter simplement si un jeu de données contient ces artefacts, Poliak *et al.* [37] ont entraîné un modèle en considérant seulement les hypothèses comme séquence d'entrée. Si le modèle est capable d'obtenir un bon score lors de l'évaluation, c'est que les hypothèses contiennent des indices sur lesquels le modèle peut s'appuyer pour « deviner » le bon label. [Jullien *et al.*, 2023a] observent des résultats similaires pour NLI4CT. Ces artefacts sont en majorité dus à l'utilisation d'annotateurs peu entraînés (par exemple recrutés via Amazon Mechanical Turk) lors de la création des exemples d'entraînement. La solution serait d'employer des annotateurs hautement qualifiés et de s'assurer de la diversité des exemples générés. Cependant cela représente un certain coût financier. De ce fait, les jeux de données de bonne qualité sont généralement de taille restreinte. Cette difficulté à recruter des annotateurs qualifiés a aussi pour conséquence de limiter la diversité linguistique des jeux de données à disposition. En effet, à l'heure où nous écrivons cet article, aucun jeu de données en français, en espagnol ou encore en italien n'est encore disponible pour l'ILN dans le domaine clinique.

Cela ouvre de nombreuses perspectives pour des travaux futurs aussi bien en général que pour le domaine clinique. Limiter les artefacts dans les jeux de données pourrait permettre de mieux appréhender la tâche. Le développement de jeux de données dans d'autres langues permettrait également une mise en lumière d'études cliniques à une échelle plus locale. Pour une tâche donnée, les méthodes basées sur des règles ou réalisant un simple entraînement sont souvent surpassées par des systèmes plus complexes utilisant par exemple l'apprentissage multi-tâche ou détectant tout d'abord les indices dans les séquences d'entrée pour ensuite appliquer la tâche d'ILN. Certaines des approches utilisant des LLM, notamment en affinant ces derniers sur des instructions, ont obtenu des résultats comparables à ceux obtenus avec des méthodes classiques. Cependant le fonctionnement de ces modèles reste plutôt opaque. Les réponses générées par ce type de modèles restent encore une piste à explorer. Certains travaux comme le jeux de données e-SNLI [7] ou encore INTERACTION [47] essayent d'évaluer et d'expliquer les justifications générées par les LLMs.

5 Conclusion et travaux futurs

Dans cet article nous avons présenté une revue de la littérature récente (de 2022 à 2023) de l'inférence en langue naturelle focalisée sur les essais cliniques. En suivant la mé-

thode PRISMA, nous avons récolté 1921 articles et après sélection nous en avons étudié 74. Nous avons relevé les différents modèles, jeux de données et approches présentés dans la littérature, à la fois pour le domaine général et clinique. Nous avons ensuite présenté quels bénéfices l'ILN pourrait apporter aux essais cliniques. Enfin nous avons présenté les différents défis et perspectives.

Nous constatons que la tâche d'ILN pour les essais cliniques est en plein essor avec l'organisation d'une campagne d'évaluation où des méthodes diverses et nombreuses ont été proposées. Des défis restent cependant à relever, comme la présence de biais dans les jeux de données, le manque de données ou le peu de diversité linguistique. Quelques études appliquent l'ILN au domaine clinique en l'utilisant comme auxiliaire pour aborder d'autres tâches de TAL (réponse à des questions, extraction de relations) ou dans des systèmes plus finalisés à destination des professionnels de santé comme des systèmes d'aide à la décision ou de sélection de patients pour des essais cliniques.

Une suite consistera à compléter cette étude avec les travaux antérieurs à 2022, que nous avons aussi collectés.

Remerciements

Ces travaux ont bénéficié du financement CNRS 80|PRIME.

Références

[1] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical BERT embeddings. *CoRR*, abs/1904.03323, 2019.

[2] Jun Bai, Chuantao Yin, Zimeng Wu, Jianfei Zhang, Yanmeng Wang, Guanyi Jia, Wenge Rong, and Zhang Xiong. Improving biomedical reqa with consistent nli-transfer and post-whitening. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(3) :1864–1875, May 2023.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert : A pretrained language model for scientific text. In *EMNLP*. Association for Computational Linguistics, 2019.

[4] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer : The long-document transformer. *CoRR*, abs/2004.05150, 2020.

[5] Rajeshree Bora-Kathariya and Yashodhara Haribhakta. Natural language inference as an evaluation measure for abstractive summarization. In *2018 4th International Conference for Convergence in Technology (I2CT)*, pages 1–4, 2018.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[7] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli : Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[8] Cemil Cengiz and Deniz Yuret. Joint training with semantic role labeling for better generalization in natural language inference. In Spandana Gella, Johannes Welbl, Marek Rei, Fabio Petroni, Patrick S. H. Lewis, Emma Strubell, Min Joon Seo, and Hannaneh Hajishirzi, editors, *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 78–88, Online, July 2020. Association for Computational Linguistics.

[9] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[10] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli : Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[11] Claudio Crema, Giuseppe Attardi, Daniele Sartiano, and Alberto Redolfi. Natural language processing in clinical neuroscience and psychiatry : A review. *Frontiers in Psychiatry*, 13, Sep 2022.

[12] Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computatio-*

nal Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] Houssein Dhayne and Rima Kilany. Using embedding-based metrics to expedite patients recruitment process for clinical trials. In *International Conference on Big Data and Cyber-Security Intelligence*, 2019.

[15] Yanjun Gao, Dmitriy Dligach, Leslie Christensen, Samuel Tesch, Ryan Laffin, Dongfang Xu, Timothy Miller, Ozlem Uzuner, Matthew M Churpek, and Majid Afshar. A scoping review of publicly available language tasks in clinical natural language processing. *Journal of the American Medical Informatics Association*, 29(10) :1797–1806, Aug 2022.

[16] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1), oct 2021.

[17] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[18] Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. A multilingual benchmark for probing negation-awareness with minimal pairs. In Arianna Bisazza and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online, November 2021. Association for Computational Linguistics.

[19] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta : Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654, 2020.

[20] Yongqun He, Sirarat Sarntivijai, Yu Lin, Zuoshuang Xiang, Abra Guo, Shelley Zhang, Desikan Jagannathan, Luca Toldo, Cui Tao, and Barry Smith. Oae : The ontology of adverse events. *Journal of Biomedical Semantics*, 5(1) :29, 2014.

[21] Christine Herlihy and Rachel Rudinger. MedNLI is not immune : Natural language inference artifacts in the clinical domain. *CoRR*, abs/2106.01491 :1020–1027, August 2021.

[22] Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. An analysis of natural language inference benchmarks through the lens of negation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online, November 2020. Association for Computational Linguistics.

[23] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert : Modeling clinical notes and predicting hospital readmission, 2020.

[24] Betina Idnay, Caitlin Dreisbach, Chunhua Weng, and Rebecca Schnall. A systematic review on natural language processing systems for eligibility prescreening in clinical research. *Journal of the American Medical Informatics Association*, 29(1) :197–206, Nov 2021.

[25] O. T. Inan, P. Tenaerts, S. A. Prindiville, H. R. Reynolds, D. S. Dizon, K. Cooper-Arnold, M. Turakhia, M. J. Pletcher, K. L. Preston, H. M. Krumholz, and et al. Digitizing clinical trials. *npj Digital Medicine*, 3(1), Jul 2020.

[26] Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are natural language inference models IMPPRESsive ? Learning IMPlicature and PRESupposition. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online, July 2020. Association for Computational Linguistics.

[27] Alice Kwak, Jacob Israelsen, Clayton Morrison, Derek Bambauer, and Mihai Surdeanu. Validity assessment of legal will statements as natural language inference. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics : EMNLP 2022*, pages 6047–6056, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[28] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC : Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10 :163–177, 2022.

[29] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4) :1234–1240, 09 2019.

[30] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta : A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[32] Bill MacCartney and Christopher D. Manning. An extended model of natural logic. In Harry Bunt, editor, *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands, January 2009. Association for Computational Linguistics.

[33] George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. UmlsBERT : Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1744–1753, Online, June 2021. Association for Computational Linguistics.

[34] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The prisma 2020 statement : an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021.

[35] Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. Scifive : a text-to-text transformer model for biomedical literature. *CoRR*, abs/2106.03598, 2021.

[36] Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. Natural language inference prompts for zero-shot emotion classification in text across corpora. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[37] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In Malvina Nissim, Jonathan Berant, and Alessandro Lenci, editors, *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[38] I Made Suwija Putra, Daniel Siahaan, and Ahmad Saikhu. Recognizing textual entailment : A review of resources, approaches, applications, and challenges. *ICT Express*, 10(1) :132–155, 2024.

[39] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020.

[41] Nils Reimers and Iryna Gurevych. Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[42] Carmen Rosa, Lisa A. Marsch, Erin L. Winstanley, Meg Brunner, and Aimee N.C. Campbell. Using digital technologies in clinical trials : Current and future applications. *Contemporary Clinical Trials*, 100 :106219, Jan 2021.

[43] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT : smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

[44] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng, and Luo Si. Structbert : Incorporating language structures into pre-training for deep language understanding. In *International Conference on Learning Representations*, 2020.

[45] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022.

[46] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert : Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*, 2022.

[47] Jialin Yu, Alexandra I. Cristea, Anoushka Harit, Zhongtian Sun, Olanrewaju Tahir Aduragba, Lei Shi, and Noura Al Moubayed. Interaction : A generative xai framework for natural language inference explanations, 2022.

[48] Xingyao Zhang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. Deepenroll : Patient-trial matching with

deep embedding and entailment prediction. In *Proceedings of The Web Conference 2020*, WWW '20, page 1029–1037, New York, NY, USA, 2020. Association for Computing Machinery.

Études incluses

[Aggarwal *et al.*, 2022] AGGARWAL, D., GUPTA, V. et KUNCHUKUTTAN, A. (2022). IndicXNLI : Evaluating multilingual inference for Indian languages. *In* GOLDBERG, Y., KOZAREVA, Z. et ZHANG, Y., éditeurs : *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10994–11006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Al Jallad et Ghneim, 2023] AL JALLAD, K. et GHNEIM, N. (2023). ArNLI : Arabic natural language inference entailment and contradiction detection. *Computer Science*, 24(2).

[Alameldin et Williamson, 2023] ALAMELDIN, A. et WILLIAMSON, A. (2023). Clemson NLP at SemEval-2023 task 7 : Applying GatorTron to multi-evidence clinical NLI. *In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1598–1602, Toronto, Canada. Association for Computational Linguistics.

[Armstrong *et al.*, 2022] ARMSTRONG, R., HEWITT, J. et MANNING, C. D. (2022). Jampatoisnli : A jamaican patois natural language inference dataset. *CoRR*, abs/2212.03419:5307–5320.

[Asael *et al.*, 2022] ASAEL, D., ZIEGLER, Z. et BELINKOV, Y. (2022). A generative approach for mitigating structural biases in natural language inference. *In* NASTASE, V., PAVLICK, E., PILEHVAR, M. T., CAMACHO-COLLADOS, J. et RAGANATO, A., éditeurs : *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 186–199, Seattle, Washington. Association for Computational Linguistics.

[Bastan *et al.*, 2022] BASTAN, M., SURDEANU, M. et BALASUBRAMANIAN, N. (2022). BioNLI : Generating a biomedical NLI dataset using lexico-semantic constraints for adversarial examples. *In* GOLDBERG, Y., KOZAREVA, Z. et ZHANG, Y., éditeurs : *Findings of the Association for Computational Linguistics : EMNLP 2022*, pages 5093–5104, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Bevan *et al.*, 2023] BEVAN, R., TURBITT, O. et ABOSHOKOR, M. (2023). MDC at SemEval-2023 task 7 : Fine-tuning transformers for textual entailment prediction and evidence retrieval in clinical trials. *In* OJHA, A. K., DOGRUÖZ, A. S., MARTINO, G. D. S., MADABUSHI, H. T., KUMAR, R. et SARTORI, E., éditeurs : *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1287–1292, Toronto, Canada. Association for Computational Linguistics.

[Chen *et al.*, 2023] CHEN, W., WEI, S., WEI, Z. et HUANG, X. (2023). KNSE : A knowledge-aware natural language inference framework for dialogue symptom status recognition. *In* ROGERS, A., BOYD-GRABER, J. L. et OKAZAKI, N., éditeurs : *Findings of the Association for Computational Linguistics : ACL 2023*, pages 10278–10286, Toronto, Canada. Association for Computational Linguistics.

[Conceição *et al.*, 2023] CONCEIÇÃO, S. I. R., F. SOUSA, D., SILVESTRE, P. et COUTO, F. M. (2023). lasigeBioTM at SemEval-2023 task 7 : Improving natural language inference baseline systems with domain ontologies. *In* OJHA, A. K., DOĞRUÖZ, A. S., DA SAN MARTINO, G., TAYYAR MADABUSHI, H., KUMAR, R. et SARTORI, E., éditeurs : *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 10–15, Toronto, Canada. Association for Computational Linguistics.

[Corrêa Dias *et al.*, 2023] CORRÊA DIAS, A., DIAS, F., MOREIRA, H., MOREIRA, V. et COMBA, J. L. (2023). Team INF-UFRGS at SemEval-2023 task 7 : Supervised contrastive learning for pair-level sentence classification and evidence retrieval. *In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 700–706, Toronto, Canada. Association for Computational Linguistics.

[Feng *et al.*, 2023] FENG, C., WANG, J. et ZHANG, X. (2023). YNU-HPCC at SemEval-2023 task7 : Multi-evidence natural language inference for clinical trial data based a BioBERT model. *In* OJHA, A. K., DOGRUÖZ, A. S., MARTINO, G. D. S., MADABUSHI, H. T., KUMAR, R. et SARTORI, E., éditeurs : *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 664–670, Toronto, Canada. Association for Computational Linguistics.

[Gubelmann *et al.*, 2023] GUBELMANN, R., KALOULI, A.-l., NIKLAUS, C. et HANDSCHUH, S. (2023). When truth matters - addressing pragmatic categories in natural language inference (NLI) by large language models (LLMs). *In* PALMER, A. et CAMACHO-COLLADOS, J., éditeurs : *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 24–39, Toronto, Canada. Association for Computational Linguistics.

[Huang *et al.*, 2023] HUANG, M., REN, J., LIU, L., SONG, R. et YIN, W. (2023). CPIC at SemEval-2023 task 7 : GPT2-based model for multi-evidence natural language inference for clinical trial data. *In* OJHA, A. K., DOGRUÖZ, A. S., MARTINO, G. D. S., MADABUSHI, H. T., KUMAR, R. et SARTORI, E., éditeurs : *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 397–401, Toronto, Canada. Association for Computational Linguistics.

[Huynh *et al.*, 2022] HUYNH, T. V., NGUYEN, K. V. et NGUYEN, N. L.-T. (2022). ViNLI : A Vietnamese corpus for studies on open-domain natural language infe-

rence. *In* CALZOLARI, N., HUANG, C., KIM, H., PUSTEJOVSKY, J., WANNER, L., CHOI, K., RYU, P., CHEN, H., DONATELLI, L., JI, H., KUROHASHI, S., PAGGIO, P., XUE, N., KIM, S., HAHM, Y., HE, Z., LEE, T. K., SANTUS, E., BOND, F. et NA, S., éditeurs : *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3858–3872, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

[Jullien *et al.*, 2023a] JULLIEN, M., VALENTINO, M., FROST, H., O'REGAN, P., LANDERS, D. et FREITAS, A. (2023a). NLI4CT : multi-evidence natural language inference for clinical trial reports. *CoRR*, abs/2305.03598.

[Jullien *et al.*, 2023b] JULLIEN, M., VALENTINO, M., FROST, H., O'REGAN, P., LANDERS, D. et FREITAS, A. (2023b). SemEval-2023 task 7 : Multi-evidence natural language inference for clinical trial data. *In* OJHA, A. K., DOGRUÖZ, A. S., MARTINO, G. D. S., MADABUSHI, H. T., KUMAR, R. et SARTORI, E., éditeurs : *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, volume abs/2305.02993, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

[Kanakarajan et Sankarasubbu, 2023] KANAKARAJAN, K. R. et SANKARASUBBU, M. (2023). Saama AI research at SemEval-2023 task 7 : Exploring the capabilities of flan-t5 for multi-evidence natural language inference in clinical trial data. *In* OJHA, A. K., DOGRUÖZ, A. S., MARTINO, G. D. S., MADABUSHI, H. T., KUMAR, R. et SARTORI, E., éditeurs : *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 995–1003, Toronto, Canada. Association for Computational Linguistics.

[Kann *et al.*, 2022] KANN, K., EBRAHIMI, A., MAGER, M., ONCEVAY, A., ORTEGA, J. E., RIOS, A., FAN, A., GUTIERREZ-VASQUES, X., CHIRUZZO, L., LUGO, G. A. G., RAMOS, R., RUÍZ, I. V. M., MAGER, E., CHAUDHARY, V., NEUBIG, G., PALMER, A., SOLANO, R. A. C. et VU, N. T. (2022). Americasnli : Machine translation and natural language inference systems for indigenous languages of the americas. *Frontiers Artif. Intell.*, 5:995667.

[Kumar Upadhyay et Kumar Upadhya, 2023] KUMAR UPADHYAY, A. et KUMAR UPADHYA, H. (2023). Xnli 2.0 : Improving xnli dataset and performance on cross lingual understanding (xlu). *In 2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pages 1–6.

[Li *et al.*, 2023] LI, S., HU, X., LIN, L., LIU, A., WEN, L. et YU, P. S. (2023). A multi-level supervised contrastive learning framework for low-resource natural language inference. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:1771–1783.

[Li *et al.*, 2022] LI, S., HU, X., LIN, L. et WEN, L. (2022). Pair-level supervised contrastive learning for natural language inference. *In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, volume abs/2201.10927, pages 8237–8241. IEEE.

[Liu *et al.*, 2022] LIU, A., SWAYAMDIPTA, S., SMITH, N. A. et CHOI, Y. (2022). WANLI : Worker and AI collaboration for natural language inference dataset creation. *In* GOLDBERG, Y., KOZAREVA, Z. et ZHANG, Y., éditeurs : *Findings of the Association for Computational Linguistics : EMNLP 2022*, volume abs/2201.05955, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Mersinias et Valvis, 2022] MERSINIAS, M. et VALVIS, P. (2022). Mitigating dataset artifacts in natural language inference through automatic contextual data augmentation and learning optimization. *In* CALZOLARI, N., BÉCHET, F., BLACHE, P., CHOUKRI, K., CIERI, C., DECLERCK, T., GOGGI, S., ISAHARA, H., MAEGAARD, B., MARIANI, J., MAZO, H., ODIJK, J. et PIPERIDIS, S., éditeurs : *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 427–435, Marseille, France. European Language Resources Association.

[Noor Mohamed et Srinivasan, 2023] NOOR MOHAMED, S. S. et SRINIVASAN, K. (2023). SSNSheerinKavitha at SemEval-2023 task 7 : Semantic rule based label prediction using TF-IDF and BM25 techniques. *In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 950–957, Toronto, Canada. Association for Computational Linguistics.

[Pahwa et Pahwa, 2023] PAHWA, B. et PAHWA, B. (2023). BpHigh at SemEval-2023 task 7 : Can fine-tuned cross-encoders outperform GPT-3.5 in NLI tasks on clinical trial data ? *In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1936–1944, Toronto, Canada. Association for Computational Linguistics.

[Rajamanickam et Rajaraman, 2023] RAJAMANICKAM, S. et RAJARAMAN, K. (2023). I2R at SemEval-2023 task 7 : Explanations-driven ensemble approach for natural language inference over clinical trial data. *In* OJHA, A. K., DOGRUÖZ, A. S., MARTINO, G. D. S., MADABUSHI, H. T., KUMAR, R. et SARTORI, E., éditeurs : *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1630–1635, Toronto, Canada. Association for Computational Linguistics.

[Sadat et Caragea, 2022] SADAT, M. et CARAGEA, C. (2022). Scinli : A corpus for natural language inference on scientific text. *In* MURESAN, S., NAKOV, P. et VILLAVICENCIO, A., éditeurs : *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, volume abs/2203.06728, pages 7399–7409. Association for Computational Linguistics.

[Saxon *et al.*, 2023] SAXON, M., WANG, X., XU, W. et WANG, W. Y. (2023). PECO : Examining single sen-

tence label leakage in natural language inference datasets through progressive evaluation of cluster outliers. *In* VLACHOS, A. et AUGENSTEIN, I., éditeurs : *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3061–3074, Dubrovnik, Croatia. Association for Computational Linguistics.

[Sosa *et al.*, 2023] SOSA, D., SURESH, M., POTTS, C. et ALTMAN, R. (2023). Detecting contradictory COVID-19 drug efficacy claims from biomedical literature. *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 694–713, Toronto, Canada. Association for Computational Linguistics.

[Takehana *et al.*, 2023] TAKEHANA, C., LIM, D., KURTULUS, E., IYER, R., TANIMURA, E., AGGARWAL, P., CANTILLON, M., YU, A., KHAN, S. et CHI, N. (2023). Stanford MLab at SemEval 2023 task 7 : Neural methods for clinical trial report NLI. *In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1769–1775, Toronto, Canada. Association for Computational Linguistics.

[Truong *et al.*, 2022] TRUONG, T. H., OTMAKHOVA, Y., BALDWIN, T., COHN, T., LAU, J. H. et VERSPOOR, K. (2022). Not another negation benchmark : The NaN-NLI test suite for sub-clausal negation. *In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 883–894, Online only. Association for Computational Linguistics.

[Vassileva *et al.*, 2023] VASSILEVA, S., GRAZHDANSKI, G., BOYTCHEVA, S. et KOYCHEV, I. (2023). FMI-SU at SemEval-2023 task 7 : Two-level entailment classification of clinical trials enhanced by contextual data augmentation. *In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1454–1462, Toronto, Canada. Association for Computational Linguistics.

[Vladika et Matthes, 2023] VLADIKA, J. et MATTHES, F. (2023). Sebis at SemEval-2023 task 7 : A joint system for natural language inference and evidence retrieval from clinical trial reports. *In* OJHA, A. K., DOGRUÖZ, A. S., MARTINO, G. D. S., MADABUSHI, H. T., KUMAR, R. et SARTORI, E., éditeurs : *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, volume abs/2304.13180, pages 1863–1870, Toronto, Canada. Association for Computational Linguistics.

[Volosincu *et al.*, 2023] VOLOSINCU, M., LUPU, C., TRANDABAT, D. et GIFU, D. (2023). FII SMART at SemEval 2023 task7 : Multi-evidence natural language inference for clinical trial data. *In* OJHA, A. K., DOGRUÖZ, A. S., MARTINO, G. D. S., MADABUSHI, H. T., KUMAR, R. et SARTORI, E., éditeurs : *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 212–220, Toronto, Canada. Association for Computational Linguistics.

[Wang *et al.*, 2023] WANG, W., XU, B., FANG, T., ZHANG, L. et SONG, Y. (2023). KnowComp at SemEval-2023 task 7 : Fine-tuning pre-trained language models for clinical trial entailment identification. *In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1–9, Toronto, Canada. Association for Computational Linguistics.

[Widiana *et al.*, 2022] WIDIANA, P. G. A. T., PURWARIANTI, A. et RUSKANDA, F. Z. (2022). Developing covid-19 information validation system using natural language inference. *In 2022 9th International Conference on Advanced Informatics : Concepts, Theory and Applications (ICAICTA)*, pages 1–6.

[Zhao *et al.*, 2023] ZHAO, X., ZHANG, M., MA, M., SU, C., LIU, Y., WANG, M., QIAO, X., GUO, J., LI, Y. et MA, W. (2023). HW-TSC at SemEval-2023 task 7 : Exploring the natural language inference capabilities of ChatGPT and pre-trained language model for clinical trial. *In* OJHA, A. K., DOGRUÖZ, A. S., MARTINO, G. D. S., MADABUSHI, H. T., KUMAR, R. et SARTORI, E., éditeurs : *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1603–1608, Toronto, Canada. Association for Computational Linguistics.

[Zhou *et al.*, 2023] ZHOU, Y., JIN, Z., LI, M., LI, M., LIU, X., YOU, X. et WU, J. (2023). THiFLY research at SemEval-2023 task 7 : A multi-granularity system for CTR-based textual entailment and evidence retrieval. *In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1681–1690, Toronto, Canada. Association for Computational Linguistics.

Tableaux des études incluses

Title	Authors	Year
A Generative Approach for Mitigating Structural Biases in Natural Language Inference	Asael *et al.*	2022
A large language model for electronic health records	Yang *et al.*	2022
A Linguistic Investigation of Machine Learning based Contradiction Detection Models : An Empirical Analysis and Future Perspectives	Pielka *et al.*	2022
AmericasNLI : Machine translation and natural language inference systems for Indigenous languages of the Americas	Kann *et al.*	2022
BioNLI : Generating a Biomedical NLI Dataset Using Lexico-semantic Constraints for Adversarial Examples	Bastan *et al.*	2022
Building a Vietnamese Dataset for Natural Language Inference Models	Nguyen and Nguyen	2022
Chinese Textual Entailment Recognition Model Based on Contextual Feature Extraction	Hu and Sui	2022
Developing COVID-19 Information validation system Using Natural language inference	Widiana *et al.*	2022
Diff-Explainer : Differentiable Convex Optimization for Explainable Multi-hop Inference	Thayaparan *et al.*	2022
Embarrassingly Simple Performance Prediction for Abductive Natural Language Inference	Kadiķis *et al.*	2022
Enhancing Cross-lingual Natural Language Inference by Prompt-learning from Cross-lingual Templates	Qi *et al.*	2022
Enhancing Natural Language Inference of Cross-lingual N-shot Transfer with Multilingual Data	Tseng and Lin	2022
FaiRR : Faithful and Robust Deductive Reasoning over Natural Language	Sanyal *et al.*	2022
Feature Fusion Transformer Network for Natural Language Inference	Sun and Yan	2022
IndicXNLI : Evaluating Multilingual Inference for Indian Languages	Aggarwal *et al.*	2022
INTERACTION : A Generative XAI Framework for Natural Language Inference Explanations	Yu *et al.*	2022
Investigating Reasons for Disagreement in Natural Language Inference	Jiang and Marneffe	2022
JamPatoisNLI : A Jamaican Patois Natural Language Inference Dataset	Armstrong *et al.*	2022
Mitigating Dataset Artifacts in Natural Language Inference Through Automatic Contextual Data Augmentation and Learning Optimization	Mersinias and Valvis	2022
Natural Language Inference for Arabic using Recurrent Neural Network and Word Embedding	Bensghaier *et al.*	2022
Natural language inference model for the Vietnamese language with machine learning algorithms : a view from “Truyen Kieu”	Mai Trang and Phung	2022
Network based on the synergy of knowledge and context for natural language inference	Wu and Huang	2022
Not another Negation Benchmark : The NaN-NLI Test Suite for Sub-clausal Negation	Truong *et al.*	2022
Pair-Level Supervised Contrastive Learning for Natural Language Inference	Li *et al.*	2022
Persian Natural Language Inference : A Meta-learning Approach	Soudani et al	2022
R2F : A General Retrieval, Reading and Fusion Framework for Document-level Natural Language Inference	Wang *et al.*	2022
Research on Judgment Reasoning Using Natural Language Inference in Chinese Medical Texts	Li and Kong	2022
SciNLI : A Corpus for Natural Language Inference on Scientific Text	Sadat and Caragea	2022
Semantic Reasoning with NLI for Assertion Detection in Medical Text	Du *et al.*	2022
SILT : Efficient transformer training for inter-lingual inference	Huertas-Tato *et al.*	2022
The Chinese Causative-Passive Homonymy Disambiguation : an adversarial Dataset for NLI and a Probing Task	Xu and Markert	2022
ViNLI : A Vietnamese Corpus for Studies on Open-Domain Natural Language Inference	Huynh *et al.*	2022
WANLI : Worker and AI Collaboration for Natural Language Inference Dataset Creation	Liu *et al.*	2022

TABLE 1 – Liste des études parues en 2022 incluses dans notre revue de la littérature

Title	Authors	Year
A Multi-Level Supervised Contrastive Learning Framework for Low-Resource Natural Language Inference	Li *et al.*	2023
ArNLI : Arabic Natural Language Inference for Entailment and Contradiction Detection	Al Jallad and Gheneim	2023
A semantics-aware approach for multilingual natural language inference	Le-Hong and Cambria	2023
Bf3R at SemEval-2023 Task 7 : a text similarity model for textual entailment and evidence retrieval in clinical trials and animal studies	Neves	2023
BpHigh at SemEval-2023 Task 7 : Can Fine-tuned Cross-encoders Outperform GPT-3.5 in NLI Tasks on Clinical Trial Data ?	Pahwa and Pahwa	2023
Clemson NLP at SemEval-2023 Task 7 : Applying GatorTron to Multi-Evidence Clinical NLI	Alameldin and Williamson	2023
CPIC at SemEval-2023 Task 7 : GPT2-Based Model for Multi-evidence Natural Language Inference for Clinical Trial Data	Huang *et al.*	2023
Detecting Contradictory COVID-19 Drug Efficacy Claims from Biomedical Literature	Sosa *et al.*	2023
Enhancing Cross-lingual Natural Language Inference by Soft Prompting with Multilingual Verbalizer	Li *et al.*	2023
Explainable Natural Language Inference via Identifying Important Rationales	Yang *et al.*	2023
FII SMART at SemEval 2023 Task7 : Multi-evidence Natural Language Inference for Clinical Trial Data	Volosincu *et al.*	2023
FMI-SU at SemEval-2023 Task 7 : Two-level Entailment Classification of Clinical Trials Enhanced by Contextual Data Augmentation	Vassileva *et al.*	2023
Generating knowledge aware explanation for natural language inference	Yang *et al.*	2023
HW-TSC at SemEval-2023 Task 7 : Exploring the Natural Language Inference Capabilities of ChatGPT and Pre-trained Language Model for Clinical Trial	Zhao *et al.*	2023
I2R at SemEval-2023 Task 7 : Explanations-driven Ensemble Approach for Natural Language Inference over Clinical Trial Data	Rajamanickam and Rajaraman	2023
Investigating Multi-source Active Learning for Natural Language Inference	Snijders *et al.*	2023
ITTC at SemEval 2023-Task 7 : Document Retrieval and Sentence Similarity for Evidence Retrieval in Clinical Trial Data	Mahendra *et al.*	2023
JUST-KM at SemEval-2023 Task 7 : Multi-evidence Natural Language Inference using Role-based Double Roberta-Large	Alissa and Abdullah	2023
KnowComp at SemEval-2023 Task 7 : Fine-tuning Pre-trained Language Models for Clinical Trial Entailment Identification	Wang *et al.*	2023
Knowledge Injection for Disease Names in Logical Inference between Japanese Clinical Texts	Murakami *et al.*	2023
KNSE : A Knowledge-aware Natural Language Inference Framework for Dialogue Symptom Status Recognition	Chen *et al.*	2023
kogito : A Commonsense Knowledge Inference Toolkit	Ismayilzada and Bosselut	2023
lasigeBioTM at SemEval-2023 Task 7 : Improving Natural Language Inference Baseline Systems with Domain Ontologies	Conceição *et al.*	2023
Leveraging Symbolic Knowledge Bases for Commonsense Natural Language Inference Using Pattern Theory	Aakur and Sarkar	2023
MDC at SemEval-2023 Task 7 : Fine-tuning Transformers for Textual Entailment Prediction and Evidence Retrieval in Clinical Trials	Bevan *et al.*	2023

TABLE 2 – Liste des études parues en 2023 incluses dans notre revue de la littérature (1/2)

Title	Authors	Year
NatLogAttack : A Framework for Attacking Natural Language Inference Models with Natural Logic	Zheng and Zhu	2023
NCUEE-NLP at SemEval-2023 Task 7 : Ensemble Biomedical LinkBERT Transformers in Multi-evidence Natural Language Inference for Clinical Trial Data	Chen *et al.*	2023
NLI4CT : Multi-Evidence Natural Language Inference for Clinical Trial Reports	Jullien *et al.*	2023
PECO : Examining Single Sentence Label Leakage in Natural Language Inference Datasets through Progressive Evaluation of Cluster Outliers	Saxon *et al.*	2023
Prompting for explanations improves Adversarial NLI. Is this true ? Yes it is true because it weakens superficial cues	Kavumba *et al.*	2023
Saama AI Research at SemEval-2023 Task 7 : Exploring the Capabilities of Flan-T5 for Multi-evidence Natural Language Inference in Clinical Trial Data	Kanakarajan and Sankarasubbu	2023
Sebis at SemEval-2023 Task 7 : A Joint System for Natural Language Inference and Evidence Retrieval from Clinical Trial Reports	Vladika and Matthes	2023
SemEval-2023 Task 7 : Multi-Evidence Natural Language Inference for Clinical Trial Data	Jullien *et al.*	2023
SSNSheerinKavitha at SemEval-2023 Task 7 : Semantic Rule Based Label Prediction Using TF-IDF and BM25 Techniques	Noor Mohamed and Srinivasan	2023
Stanford MLab at SemEval 2023 Task 7 : Neural Methods for Clinical Trial Report NLI	Takehana *et al.*	2023
Team INF-UFRGS at SemEval-2023 Task 7 : Supervised Contrastive Learning for Pair-level Sentence Classification and Evidence Retrieval	Corrêa Dias *et al.*	2023
THiFLY Research at SemEval-2023 Task 7 : A Multi-granularity System for CTR-based Textual Entailment and Evidence Retrieval	Zhou *et al.*	2023
Uncertainty-Aware Natural Language Inference with Stochastic Weight Averaging	Talman *et al.*	2023
Unsupervised Contradiction Detection using Sentence Transformations	Schumann and Gómez	2023
When Truth Matters - Addressing Pragmatic Categories in Natural Language Inference (NLI) by Large Language Models (LLMs)	Gubelmann *et al.*	2023
XNLI 2.0 : Improving XNLI dataset and performance on Cross Lingual Understanding (XLU)	Kumar Upadhyay and Kumar Upadhya	2023
YNU-HPCC at SemEval-2023 Task7 : Multi-evidence Natural Language Inference for Clinical Trial Data Based a BioBERT Model	Feng *et al.*	2023

TABLE 3 – Liste des études parues en 2023 incluses dans notre revue de la littérature (2/2)