

Inférence en langue naturelle appliquée au recrutement de patients pour les essais cliniques : le point de vue du patient

Mathilde Aguiar Pierre Zweigenbaum Nona Naderi

Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France
{prenom.nom}@lisn.fr

RÉSUMÉ

Recruter des patients pour les essais cliniques est long et complexe. Habituellement, le processus de recrutement est initié par un professionnel de santé qui propose à un patient de participer à l'essai clinique. Promouvoir les essais directement aux patients via des plateformes en ligne pourrait aider à en atteindre un plus grand nombre. Dans cette étude, nous nous intéressons au cas où le patient est l'initiateur de la démarche et veut savoir s'il est éligible à un essai clinique, tout cela en utilisant son propre langage patient. Pour déterminer si l'utilisation d'un tel langage permet tout de même au modèle de langue de déterminer l'éligibilité du patient pour l'essai clinique, nous construisons la tâche *Natural Language Inference for Patient Recruitment* (NLI4PR). Pour cela nous adaptons le jeu de données TREC 2022 Clinical Trial Track en réécrivant manuellement les profils médicaux en langage patient. Nous extrayons également les essais cliniques où les patients étaient labellisés « éligible » ou « exclu ». Nous soumettons des amorces à plusieurs grands modèles de langue, et obtenons un score F1 compris entre 56,6 et 71,8 avec le langage patient, contre 64,7 à 73,1 pour du langage médical. Nous observons que l'utilisation du langage patient ne mène qu'à une dégradation de performance relativement petite pour notre meilleur modèle. Cela suggère qu'avoir le patient en tant que point de départ du recrutement pourrait être réalisable. Nos scripts ainsi que nos jeux de données sont disponibles sur Github¹ et HuggingFace² (Aguiar *et al.*, 2025).

ABSTRACT

Am I eligible? Natural Language Inference for Clinical Trial Patient Recruitment : the Patient's Point of View

Recruiting patients to participate in clinical trials can be challenging and time-consuming. Usually, participation in a clinical trial is initiated by a healthcare professional and proposed to the patient. Promoting clinical trials directly to patients via online recruitment might help to reach them more efficiently. In this study, we address the case where a patient is initiating their own recruitment process and wants to determine whether they are eligible for a given clinical trial, using their own language to describe their medical profile. To study whether this creates difficulties in the patient-trial matching process, we design a new dataset and task, Natural Language Inference for Patient Recruitment (NLI4PR), in which patient-language profiles must be matched to clinical trials. We create it by adapting the TREC 2022 Clinical Trial Track dataset, which provides patients' medical profiles, and rephrasing them manually using patient language. We also use the associated clinical trial reports where the patients are either eligible or excluded. We prompt several open-source Large Language Models on our task and achieve from 56.5 to 71.8 of F1 score using patient language, against 64.7 to 73.1 for the same task using medical language. When using patient language, we observe only a

1. <https://github.com/CTInfer/NLI4PR>

2. <https://huggingface.co/datasets/Mathilde/NLI4PR>

small loss in performance for the best model, suggesting that having the patient as a starting point could be adopted to help recruit patients for clinical trials. The corpus and code bases are all freely available on our Github and HuggingFace repositories ([Aguilar et al., 2025](#)).

MOTS-CLÉS : Inférence en Langue Naturelle, Essais cliniques, Recrutement de patients, Grands Modèles de Langue.

KEYWORDS: Natural Language Inference, Clinical trials, Patient recruitment, Large Language Models.

ARTICLE : **Accepté à CL4Health 2025.**

Références

AGUIAR M., ZWEIGENBAUM P. & NADERI N. (2025). Am I eligible? natural language inference for clinical trial patient recruitment : the patient's point of view. In S. ANANIADOU, D. DEMNER-FUSHMAN, D. GUPTA & P. THOMPSON, Édts., *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, p. 243–259, Albuquerque, New Mexico : Association for Computational Linguistics.