Compte-rendu

# Médecins et intelligence artificielle : une alliance gagnante ?

## Mots-clés

IA, intelligence artificielle, LLM, large language model, ChatGPT
GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial

E. Goh et al., Nature Medicine, 5 février 2025

DOI: 10.1038/s41591-024-03456-y

## Introduction

**Les grands modèles de langage (Large Language Model abrégé LLM), tels que ChatGPT, regroupent un type d'intelligence artificielle spécifiquement conçu pour traiter et générer du langage humain. Les LLM ont montré des progrès remarquables ces dernières années en médecine dans le raisonnement diagnostique, si bien que plusieurs études ont mis en lumière leur supériorité par rapport aux médecins « humains » dans leur aptitude à établir des diagnostics différentiels, à expliquer leur raisonnement et à recueillir une anamnèse auprès de patients standardisés. En revanche, leur capacité de raisonnement de prise en charge — impliquant des décisions complexes telles que le choix des examens, des traitements, la balance bénéfice-risque, la prise en compte des préférences du patient, des déterminants sociaux de santé et de la prise en compte des coûts — reste encore peu explorée. Cette étude a pour objectif de déterminer si l'usage d'un LLM comme outil d'aide au raisonnement clinique améliore les performances des médecins.**

## Méthode

Étude multicentrique (États-Unis), prospective randomisée contrôlée en simple aveugle, comparant des médecins de spécialités générales répondant à une série questions complexes sur la prise en charge de vignettes cliniques, randomisés en 2 groupes : l'un avec et l'autre sans utilisation de GPT-4. Les deux groupes avaient accès aux ressources conventionnelles de pratique clinique (UpToDate, internet, etc.). Les résultats des 2 groupes ont également été comparés avec un bras LLM seul, soit sans l'intervention d'un médecin. Un score de résultat a été développé avec l'aide d'un groupe d'expert, tenant compte de la diversité de réponses acceptables, sous forme de pourcentage, sans seuil de réussite. Issue primaire : score moyen de chaque groupe. Issues secondaires : score des sous-groupes de questions (diagnostic, prise en charge, etc.), temps passé sur chaque cas.

## Résultats

**92 médecins** randomisés entre novembre 2023 et avril 2024, 73% de médecin chefs et 27% de médecins assistants. 74% spécialisés en médecine interne, 20% en médecine d'urgence et 6% de médecine de famille et communautaire. Durée médiane d'exercice dans la profession de 5.8 ans. 400 vignettes cliniques évaluées. Les médecins du groupe utilisant GPT-4 ont obtenu un score moyen significativement supérieur au groupe utilisant uniquement des ressources conventionnelles (**43% vs 35.7%, p<0.001**), avec notamment meilleurs résultats dans les sous-groupes de questions relatives aux décisions diagnostiques, à la prise en charge et spécifique au contexte. GPT-4 seul a montré des résultats similaires au groupe des médecins utilisant GPT-4 (**43.7% vs 40.0%, p=0.80**), ainsi que de meilleurs résultats, quoique non significatifs, par rapport au groupe médecin n'ayant pas utilisé GPT-4 (**43.7% vs 35.7%, p=0.074**). Les médecins du groupe GPT-4 ont passé en moyenne plus de temps sur chaque cas (**801.5 vs 690.2 sec., p=0.022**). L'analyse des risques était similaire entre les 2 groupes.

# Discussion

Les résultats de cette étude soutiennent l'utilisation de LLM comme complément utile au raisonnement du clinicien, allant même jusqu'à suggérer un potentiel d'application autonome pour certains scénarios cliniques. Le groupe de médecins utilisant GPT-4 ayant passé plus de temps sur chaque vignette, on peut se demander si le LLM incite à ralentir et réfléchir davantage ou s'il améliore réellement le raisonnement. La limite principale de cette étude est qu'elle repose sur des vignettes cliniques et non sur des situations réelles.

# Conclusion

Les **LLM sont prometteurs pour le soutien à la décision du médecin**, y compris dans des tâches aussi complexes que le raisonnement clinique et leur utilité sur des cas cliniques réels reste encore à valider.

| Date de publication | Auteurs |
|---|---|
| 12.05.2025 | |