
L'IA peut-elle remplacer le raisonnement clinique d'un médecin ?

BERTIN Hugo

Résumé

Cette synthèse interroge sur le rôle de l'intelligence artificielle (IA) dans le raisonnement médical et ses conséquences pour la pratique clinique. Face à des outils algorithmiques de plus en plus efficaces, nous pouvons nous questionner sur leur intégration dans la pratique médicale. Si certaines études montrent que ces outils peuvent égaler voire dépasser la précision diagnostique de médecins, nous constatons encore des limites. Ces observations invitent à repenser le rôle de l'IA comme un outil collaboratif, susceptible d'améliorer l'efficacité des soins tout en gardant le jugement humain du médecin.

I. Introduction

L'intelligence artificielle fait référence à l'ensemble de procédés informatiques destinés à imiter certaines opérations humaines, telles que le raisonnement, l'acquisition de connaissances ou la compréhension du langage. Elle repose sur des systèmes capables d'analyser de grandes quantités de données pour en extraire des modèles et de pronostiquer des décisions.

L'intelligence artificielle a connu ces dernières années une croissance rapide dans le secteur médical. Parmi les dernières technologies, on retrouve les grands modèles de langage (LLM), capables de gérer d'importantes quantités de données. Ces modèles suscitent un intérêt croissant

en médecine en raison de leur capacité à traiter et à synthétiser de grandes quantités d'informations.

Le raisonnement clinique s'appuie sur la collecte et l'analyse des données du patient ou encore la décision. L'apparition de LLM capables d'atteindre des résultats similaires, voire meilleurs, à ceux des médecins dans certaines recherches soulève donc la question de leur aptitude à reproduire ce type de raisonnement.

Cependant, les résultats obtenus dans des cadres expérimentaux ne permettent pas forcément d'affirmer qu'un médecin peut être remplacé. En effet, le raisonnement clinique ne se limite pas seulement à

l'examen de données médicales, mais englobe également des aspects contextuels, relationnels et éthiques qu'un robot ne peut avoir ou d'une manière limitée.

Cette synthèse vise à examiner les informations récentes sur les compétences de l'IA et notamment des LLM en matière de raisonnement médical. À la suite de la présentation des performances diagnostiques, nous analyserons les limites actuelles de ces modèles avant d'aborder les perspectives d'une coopération entre le professionnel de santé et ces nouveaux outils.

II. Les performances diagnostiques rapprochant l'IA d'un raisonnement médical

L'évaluation des capacités des LLM constitue l'un des principaux axes de recherche développés au cours des dernières années. Plusieurs études ont cherché à déterminer si ces modèles étaient capables d'atteindre un niveau de performance comparable à celui des médecins.

Les essais cliniques utilisent généralement un comparatif via des vignettes cliniques qui sont des cas cliniques et histoires cliniques amenant à des questions par un contexte, des imageries ou encore des biologies. Les médecins, quel que soit la

spécialité, vont être évalués sur plusieurs paramètres. Ces essais peuvent être randomisés ou non randomisés. Cette différence se place sur la répartition : dans les essais randomisés, les participants sont répartis au hasard entre les groupes (limitant les biais). Cette randomisation permet de répartir de façon équilibrée des caractéristiques des participants comme l'âge ou le sexe. Dans les essais non randomisés, les participants sont répartis selon un choix. Cela peut donc impliquer des différences initiales entre les groupes et par conséquent augmenter le risque des biais.

Dans un essai clinique, un LLM a atteint une précision de 80 % sur le premier diagnostic différentiel, contre 50 à 70 % pour les médecins [3]. Concernant les deux premiers diagnostics différentiels combinés, l'interface a obtenu un taux de 100 %, contre 70 à 90 % pour les médecins (Fig. 1). Au-delà de la précision, l'interface a réalisé les consultations en moins de temps mais surtout avec un coût relativement faible vis-à-vis des médecins. Malgré cela, la satisfaction des patients est légèrement inférieure aux médecins mais celle-ci reste néanmoins comparable car

cette différence est relativement faible comparé aux autres résultats.

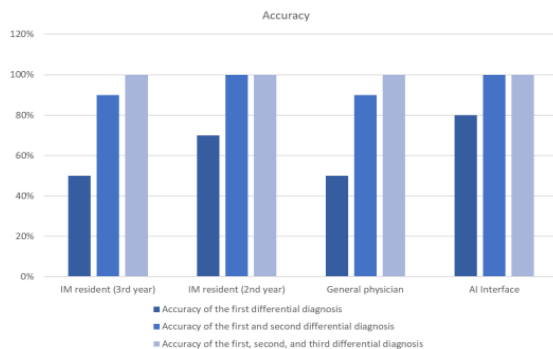


Fig 1 : Graphique montrant la comparaison des résultats de diagnostics entre une IA et des médecins

Des résultats similaires sont observés en dermatologie, où l'IA montre encore des capacités de reconnaissance lésionnelle comparables à celles de dermatologues [1], confirmant que ces performances ne sont pas propres à un seul domaine médical mais à toutes les spécialités médicales, comme nous avons pu le voir dans les exemples précédents.

Certains modèles sont donc d'excellents analystes pour trouver un diagnostic juste. Cependant, nous pouvons aussi constater une performance de qualité concernant leur raisonnement pour arriver au résultat. Le résultat est effectivement un élément à prendre en compte, mais la façon dont nous y sommes arrivés est autant voire plus pertinent. Ce qui distingue un bon médecin est sa compétence à émettre des hypothèses et à justifier ses choix en les hiérarchisant selon un contexte.

Comme évoqué précédemment, ces modèles de données sont enrichis par des mises à jour constantes pour améliorer leur précision et ajouter de nouvelles possibilités. Cette capacité d'évolution reste un avantage non négligeable face à un domaine où les avancées et les connaissances sont en constante progression. La médecine est en effet une discipline en perpétuelle évolution : des traitements sont remis en question, de nouvelles pathologies sont découvertes. C'est donc pour cela que le médecin doit lui aussi se mettre à jour de toutes les nouveautés concernant ce domaine, tout comme les LLM le font avec des mises à jour.

La qualité de raisonnement d'un LLM est notamment évaluée par le score R-IDEA (*Revised Illness Script Explication and Diagnostic Accuracy*). Cet outil s'appuie sur 4 critères principaux : l'identification du problème par l'interprétation des données, la hiérarchisation des diagnostics différentiels, la justification du diagnostic principal et enfin les diagnostics à ne pas manquer, notamment les plus dangereux. Ce score est noté sur 10, plus il est élevé, plus le raisonnement est considéré comme rigoureux et complet. C'est donc un outil qui ne mesure pas si le bon diagnostic a été trouvé, mais comment le modèle a raisonné pour y arriver. On retrouve notamment

l'utilisation de cet outil dans plusieurs études.

Nous pouvons d'abord constater cela par l'utilisation de ce score dans une étude comparant le raisonnement entre une IA, des médecins et des internes. Le chatbot possède les meilleurs résultats avec notamment un score médian de 10 là où les médecins et internes ont une note inférieure (respectivement 9 et 8) [7]. De plus, les différentes caractéristiques de ce score ont été regroupés dans un tableau (Fig. 2), pointant les points forts mais aussi les limites (cf. III) de ces systèmes comme inexactitudes correspondant à 13.8% des résultats de l'IA.

Outcome	All (N = 232) ^{a,b}	Chatbot (n = 80)	Attending physician (n = 80) ^a	Resident (n = 72)
R-IDEA score				
Median (IQR)	9 (7-10)	10 (9-10)	9 (6-10)	8 (4-9)
Score group				
Low: 0-7	90 (38.8)	4 (5.0)	29 (36.3)	33 (45.8)
High: 8-10	142 (61.2)	76 (95.0)	51 (63.8)	39 (54.9)
Correct clinical reasoning instances	222 (95.7)	80 (100)	79 (98.8)	63 (87.5)
Incorrect clinical reasoning instances	23 (9.9)	11 (13.8)	10 (12.5)	2 (2.8)
Diagnostic accuracy				
Median (IQR)	100 (100-100)	100 (100-100)	100 (100-100)	100 (100-100)
Score group				
Low: <75%	32 (13.8)	8 (10.0)	14 (17.5)	10 (13.9)
High: 75%-100%	200 (86.2)	72 (90.0)	66 (82.5)	62 (86.1)
Cannot-miss diagnoses included, median (IQR). % ^b	66.7 (33.3-100)	66.7 (50.0-100)	50.0 (27.1-100)	66.7 (33.3-81.2)

Fig 2 : Tableau comparant le score R-IDEA entre des médecins, des internes et une IA

Au-delà de la qualité de diagnostic et de raisonnement, nous pouvons constater un argument en faveur des LLM : leur capacité à produire des performances notables dans un délai et à un coût nettement inférieurs à

une consultation médicale. Ces avantages ne sont pas négligeables, notamment lorsqu'ils sont déployés à grande échelle comme pour une population.

Cependant, nous vivons actuellement un constat, les systèmes de santé font face à des défis démographiques : vieillissement de la population, pénurie de médecins ou encore l'augmentation du nombre de maladies chroniques. C'est dans ce contexte que l'utilisation de l'intelligence artificielle se pose. Ces outils sont disponibles en permanence, ils sont capables de traiter un grand nombre d'informations simultanément, ils n'ont aucune contrainte géographique. Contrairement à un médecin, ces systèmes ne connaissent pas la fatigue, ni les congés, ni les limites horaires donc cela montre une réelle force vis-à-vis d'un médecin.

Ces gains de temps et d'argent sont représentés dans plusieurs études. Certaines interfaces sont capables de trouver un bon diagnostic 44.6% plus rapidement qu'un médecin (environ 9 minutes contre 16 minutes) ainsi qu'un coût passant de 0.08 dollars à 4.2 dollars, soit une diminution de 98.1% [3].

Nous avons donc exposé les qualités indéniables des LLM. Certains sont même capables de dépasser les performances d'un professionnel de santé. Cependant, ces

résultats ont été obtenus dans des conditions expérimentales plus ou moins contrôlées, potentiellement éloignés de la complexité et l'imprévisibilité d'une consultation réelle. Il convient ainsi d'analyser les limites de ces systèmes pour rendre compte des améliorations possibles pour ces outils.

III. Les limites actuelles de l'IA face aux raisonnements cliniques

Si les performances diagnostiques des LLM sont indéniables dans des conditions expérimentales, leur déploiement en situation clinique réelle se heurte à des limites encore trop importantes. Celles-ci concernent aussi bien la qualité du raisonnement produit, la capacité à intégrer le contexte implicite du patient, que les effets à long terme de l'IA sur les compétences des médecins qui l'utilisent.

Malgré des résultats pertinents, les LLM ne sont pas exempts d'erreurs. Contrairement au médecin qui va signaler une incertitude et reconnaître ses limites, l'IA va créer une réponse avec assurance qu'il ait raison ou tort. En situation réelle, le médecin doit construire progressivement son raisonnement à partir d'informations parfois incomplètes ou imprécises. Ainsi, l'interrogatoire ou encore l'examen clinique qu'il récolte permet d'affiner les hypothèses probables.

Une limite importante concerne la transparence des systèmes d'IA. Certains LLM sont qualifiés de « *boîtes noires* » [4]. Ce terme est utilisé du fait que ces outils produisent des recommandations diagnostiques sans comprendre comment l'IA est arrivée à ces conclusions : quelles données ont été pertinentes ou encore ignorées. Cela pose un problème du fait que dans le domaine médical, le médecin doit pouvoir évaluer, discuter et si nécessaire contester une recommandation avant de l'intégrer à sa décision clinique. Un praticien ne peut s'appuyer sur un outil dont il ne comprend pas le raisonnement car ce dernier est primordial.

À cette non transparence, s'ajoute aussi le risque d'« *hallucinations* ». Les LLM peuvent générer des informations qui semblent plausibles mais qui sont en réalité incorrectes voire totalement inventées [5]. L'inconvénient est que dans un environnement clinique, ce type d'erreur est particulièrement dangereux pour le patient et peut entraîner des conséquences irréversibles. Ce phénomène est d'autant plus préoccupant dans des contextes de prises en charge rapide comme les soins intensifs. Ainsi, la capacité des LLM à produire des réponses structurées va paradoxalement constituer un certain facteur de risque pour le patient en

supposant une fiabilité qui n'existe finalement pas.

Au-delà des erreurs de raisonnement, une limite d'autant plus importante concerne l'incapacité des LLM à appréhender la dimension humaine et contextuelle de la rencontre médicale. La médecine ne se réduit pas à un traitement de données : elle implique une relation, une présence, et une compréhension globale du patient dans son environnement. Or, un LLM ne perçoit que ce qui lui est transmis par écrit, et reste aveugle à toutes émotions. Cette limite est d'autant plus importante lorsque l'on sait que ce relationnel peut orienter vers un diagnostic. Le niveau d'anxiété du patient, ses conditions de vie ou encore ses croyances vis-à-vis des soins orientent le raisonnement clinique et ne peuvent pas être perçus par un simple algorithme.

On distingue deux points : la prédiction qui est excellente chez l'IA et qui permet d'anticiper un résultat et le jugement qui nécessite une relation praticien-patient qui n'est pas possible chez un outil informatique actuellement [1]. On note également le risque de l'utilisation de l'IA chez les plus jeunes professionnels qui pourraient faire davantage confiance à ces outils qu'à leur faible expérience et donc potentiellement d'avoir de faux résultats.

Enfin, une dernière limite pourrait être abordée même si elle n'est pas immédiate mais plutôt sur un long terme, c'est la dégradation des compétences cliniques des médecins et surtout futurs médecins. L'introduction d'un outil de cette sorte peut modifier la façon dont le clinicien raisonne. Si l'IA prend de plus en plus de place dans la logique du professionnel, celui-ci pourrait créer une dépendance à l'outil. Malheureusement, lorsque cette aide ne sera pas disponible, on constatera une dégradation évidente du niveau de médecin qui est dangereuse pour le patient.

Ce principe est développé sous le terme « *deskilling* », donc la perte de(s) compétence(s). On peut le définir plus rigoureusement par la dégradation de compétences cliniques liée à une dépendance excessive d'outils algorithmiques comme un LLM [8]. Comme tout à l'heure, plusieurs types de « *deskilling* » sont constatés : le « *decision-making deskilling* » montrant la diminution des compétences techniques du médecin ou encore le « *moral deskilling* » représentant la dégradation des qualités humaines et relationnelles entretenues entre les deux camps. Cela montre bien que ces outils peuvent malheureusement abaisser le niveau des médecins, sur plusieurs plans et pas seulement sur les connaissances et la technique mais aussi sur le plan humain qui est primordial pour adresser un diagnostic

de qualité. Ces risques sont d'autant plus préoccupants lorsqu'ils se manifestent en situation de problème technologique comme une panne de l'IA par exemple. Le praticien se retrouve alors seul et peut avoir du mal à évoquer un diagnostic car trop habitué à l'informatique.

Ainsi, malgré des qualités factuelles, ces modèles de données connaissent des limites indéniables montrant que l'évolution de ces outils est toujours un défi dans la recherche. Cependant, une collaboration raisonnée et bien utilisée pourrait être une perspective pertinente en liant les avantages du médecin et du programme.

IV. Vers une collaboration médecin-IA

Les limites identifiées dans la partie précédente ne condamnent pas l'utilisation de l'IA en médecine. Nous constatons que ces objets ne seraient pas des outils pour nous remplacer mais nous devrions plutôt les repenser comme des aides complémentaires. Ces réseaux de données doivent pouvoir augmenter les capacités d'un médecin, en prenant en charge ce qu'elle sait faire aussi bien voire mieux que le professionnel, et en laissant le jugement humain à ce dernier.

Le défi reste maintenant l'intégration de ses outils dans la pratique clinique pour maximiser la qualité des diagnostics et de prises en charge. Sur long terme, une collaboration efficace entre ces deux éléments pourrait faire gagner du temps et de l'argent que ce soit pour le patient ou pour le médecin même si le point principal reste un diagnostic qualitatif.

Pour comprendre comment cette intégration peut se faire, il faut distinguer deux formes que l'IA peut prendre. Tout d'abord nous avons la notion d'« automatisation » qui vise à remplacer les humains par des systèmes algorithmiques. D'un autre côté nous avons l'« augmentation » qui cherche à compléter et enrichir les capacités de l'humain [1]. Dans le domaine médical, cette distinction est très importante : là où l'automatisation peut potentiellement menacer l'emploi et l'autonomie du professionnel, l'augmentation améliore la productivité en fournissant de l'aide à la décision. C'est pour cela que dans le monde médical, l'IA doit s'approcher de la définition d'une augmentation plutôt qu'une automatisation pour obtenir le meilleur résultat possible.

Ainsi, des modèles de collaboration entre un modèle informatique et un professionnel de santé ont été développés. Ces derniers permettent d'identifier les étapes où l'IA

peut intervenir en tant qu'outil d'assistance et son rôle (Fig. 3).

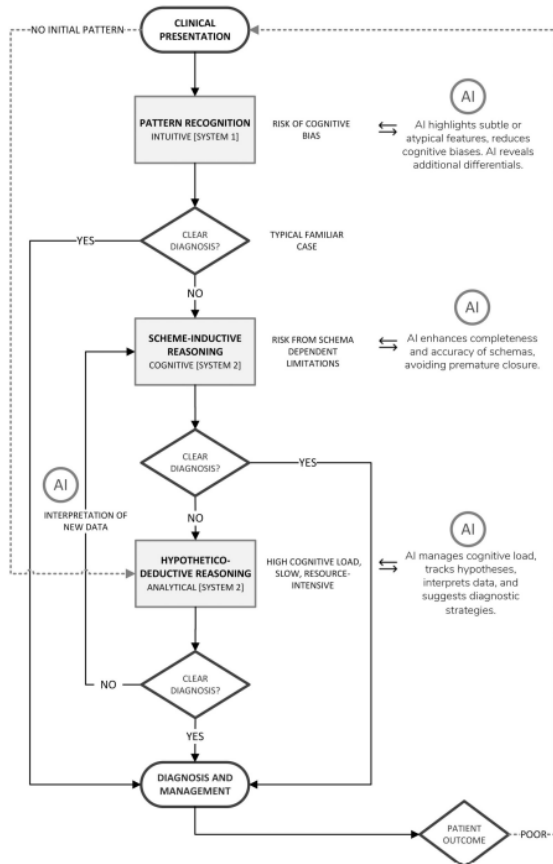


Fig 3 : Algorithme d'intégration de l'intelligence artificielle dans un raisonnement clinique

Avant la mise en place de ces éléments, il faut aussi prendre en compte certains points importants. Tout d'abord nous devons nous assurer de la bonne utilisation de l'outil par le professionnel de santé. En effet, comme dit précédemment, une mauvaise utilisation ou une utilisation excessive serait contre-productive voire dangereuse. Il faudrait donc que les médecins soient formés à l'emploi de ces produits pour en assurer une utilisation optimale [6]. Ces pratiques pourraient être intégrées dès la formation

initiale pour permettre aux futurs médecins d'avoir les compétences nécessaires pour interagir avec l'IA dans ce domaine.

Un autre point mentionné serait aussi le rapport à l'IA dans un cadre juridique. On souligne notamment la protection des données qui est le thème principal entre le droit et l'IA. Le RGPD (*Règlement Général sur la Protection des Données*) est présent pour instaurer des règles et lois autour du traitement des données permettant de définir des limites. Même avec l'existence de ce cadre, il reste encore des réglementations spécifiques à l'IA médicale nécessaires [2]. Il y'a donc la nécessité du consentement des patients. Ces derniers devraient donc être informés qu'un outil algorithmique participe à leur prise en charge.

Enfin, un autre point à aborder serait aussi l'équité en santé. En effet, les LLM sont entraînés sur des données qui peuvent sur-représenter une certaine population. Il faudrait donc s'assurer au préalable de limiter les inégalités de santé pour le déploiement d'un système comme cela à grande échelle.

V. Conclusion

Le développement de l'intelligence artificielle en médecine pose des

interrogations qui vont au-delà des simples performances techniques. Les études indiquent que l'IA, notamment les grands modèles de langage, peut obtenir un niveau de précision diagnostique impressionnant.

Cependant, ces résultats ne permettent pas d'établir une autonomie clinique de ces systèmes, il reste encore trop de limites. Ce que la science souligne principalement, c'est que l'IA ne raisonne pas comme un médecin, elle analyse, anticipe et propose. Le raisonnement clinique fait appel à bien plus qu'une simple capacité à identifier des modèles : il implique l'intuition, l'expérience, la relation avec le patient et la responsabilité éthique, des paramètres qui échappent aux algorithmes actuels.

C'est pour cette raison que le modèle le plus optimiste est celui de la coopération, où l'IA améliore les compétences du clinicien sans remplacer son discernement. Cela signifie en réalité former les professionnels de santé à une utilisation réfléchie de ces outils, et mettre en place un cadre éthique et réglementaire précis concernant la responsabilité des décisions médicales soutenues par l'IA.

Références

[1] Göndöcs, D., & Dörfler, V. (2024). AI in medical diagnosis: AI prediction & human

judgment. *Artificial Intelligence in Medicine*, 149, 102769.

[2] Prentzas, N., Kakas, A., & Pattichis, C. S. (2023). Explainable AI applications in the medical domain: a systematic review. *arXiv preprint*, arXiv:2308.05411.

[3] Park, H., Woo, C.-Y., Lim, S., Lim, S., Kwak, K., Jeong, J. Y., & Suh, C. H. (2025). Comparisons between a large language model-based real-time compound diagnostic medical AI interface and physicians for common internal medicine cases using simulated patients. *arXiv preprint*, arXiv:2505.20609.

[4] Greengrass, C. J. (2026). Transforming clinical reasoning - the role of AI in supporting human cognitive limitations. *Frontiers in Digital Health*, 7, 1715440.

[5] El Gharib K., Jundi B., Furfaro D., Abdalnour R.-E.E. (2024). AI-assisted human clinical reasoning in the ICU: beyond "to err is human". *Frontiers in Artificial Intelligence*, 7, 1506676.

[6] Goh E., Gallo R., Hom J., Strong E., Weng Y., Kerman H., Cool J.A., Kanjee Z., Parsons A.S., Ahuja N., Horvitz E., Yang D., Milstein A., Olson A.P.J., Rodman A., Chen J.H. (2024). Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open*, 7(10), e2440969.

[7] Cabral S., Restrepo D., Kanjee Z., Wilson P., Crowe B., Abdulnour R.-E., Rodman A. (2024). Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. *JAMA Internal Medicine*, 184(5), 581–583.

[8] Natali C., Marconi L., Dias Duran L.D., Cabitza F. (2025). AI-induced Deskillling in Medicine: A Mixed-Method Review and Research Agenda for Healthcare and Beyond. *Artificial Intelligence Review*, 58, 356.