Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240543

Measured Performance and Healthcare Professional Perception of Large Language Models Used as Clinical Decision Support Systems: A Scoping Review

Solène DELOURME [a,b,1], Akram REDJDAL[a], Jacques BOUAUD[a] and Brigitte SEROUSSI[a,c,d]
[a] *Sorbonne Université, Université Sorbonne Paris Nord, INSERM, Laboratoire d'Informatique Médicale et d'Ingénierie des connaissances en e-Santé, LIMICS, Paris, France*
[b] *Epita, Paris, France*
[c] *AP-HP, Hôpital Tenon, Paris,France*
[d] *APREC, Paris, France*

Abstract. The healthcare sector confronts challenges from overloaded tumor board meetings, reduced discussion durations, and care quality concerns, necessitating innovative solutions. Integrating Clinical Decision Support Systems (CDSSs) has a potential in supporting clinicians to reduce the cancer burden, but CDSSs remain poorly used in clinical practice. The emergence of OpenAI's ChatGPT in 2022 has prompted the evaluation of Large Language Models (LLMs) as potential CDSSs for diagnosis and therapeutic management. We conducted a scoping review to evaluate the utility of LLMs like ChatGPT as CDSSs in several medical specialties, particularly in oncology, and compared users' perception of LLMs with the actually measured performance of these systems.

Keywords. Clinical Decision Support Systems, Large Language Models, ChatGPT.

1. Introduction

Clinical decision support systems (CDSSs) have emerged as pivotal tools to help clinicians in their decision-making process. Despite the development of numerous CDSSs in recent years, mostly guideline-based, CDSSs remain underutilized in clinical practice, with a few only briefly adopted and not integrated into routine care [1]. With the rise of Large Language Models (LLMs), marked by the emergence of OpenAI's ChatGPT in 2022, numerous studies have been conducted to explore the performance of LLMs in diagnostic and therapeutic management of patients. LLMs can analyze medical literature and patient clinical reports to assist in diagnosis, suggest treatment recommendations and personalize care. In this paper, we carried out a scoping review to analyze studies that evaluate the actual performance of LLMs used as CDSSs with a

[1] Corresponding Author: Solène Delourme; E-mail: solene.delourme@gmail.com.

special focus on the oncology domain. Additionally, we studied users' perception of these systems and whether they would recommend to use LLMs in daily practice.

2. Methods

We conducted a scoping review, using PubMed, on articles describing LLMs as CDSSs. We focused our study on articles published in English language, between January 2010 and March 1st, 2024, involving the study of LLM proposals for the therapeutic management of patients. We removed articles that were not relevant based on titles and abstracts (see Figure 1 for the query and exclusion criteria) and selected the final studies on their full texts. Relevant references from the selected articles were also added.

Once the articles were selected, we compiled a synthesis matrix to compare them according to criteria such as publication date, the medical field covered, type of patients (either fictitious or real patients), prompt sources for LLMs, gold standard for evaluation, actual performance of ChatGPT as compared to the gold standard, and users' perception.

Users' perception was measured according to whether users would recommend to use ChatGPT as a therapeutic decision support system and was categorized into four possible responses: **Yes**, in green, when users did recommend to use ChatGPT as a decision support tool (positive feeling), **Temper**, in yellow, when users considered that LLMs were not reliable yet, but had promising potential (positive feeling), **No**, in red, when users considered that LLMs were not yet able to be used as a decision-making tool (negative feeling), and **Neutral** in gray, when no opinion was expressed (neutral feeling). Users' perceptions were manually extracted from retrieved articles, taking into account direct quotes, context, and study results.

Prompts were classified according to their source into four categories: general questions (**general**), questions related to guidelines (**guidelines**), questions about the resolution of clinical cases (**clinical cases**), and questions asked by patients (**patients**).

Evaluation metrics concentrate on ChatGPT's precision as compared to the gold standard. ChatGPT's precision was segmented into three levels: **Low, Medium, and High,** when the precision was lower than 60%, between 60% and 80%, and above 80%, resp. We added **No results** for studies lacking precision for the comparison of ChatGPT and the gold standard. Finally, a cross-evaluation was designed to analyze the link between ChatGPT's actual precision and users' perceptions of LLMs.

3. Results

From the initial 260 studies retrieved by the PubMed query, 198 articles were excluded based on titles and abstracts, 43 were excluded after the full text analysis, yielding to 19 papers to which two references were added to get at the end 21 studies (see Figure 1).

The selected articles cover 10 medical specialties (the synthesis matrix is displayed in Table 1). Twelve papers involve clinical cases, among which seven worked on fictitious patients and five worked on real patients.

Table 2 outlines the link between LLMs' performance and users' perceptions, highlighting a positive perception towards LLMs as CDSSs (7 Yes and 6 Temper) even in a few studies where the measured performance was classified as 'Medium', 'Low', or 'No result'.

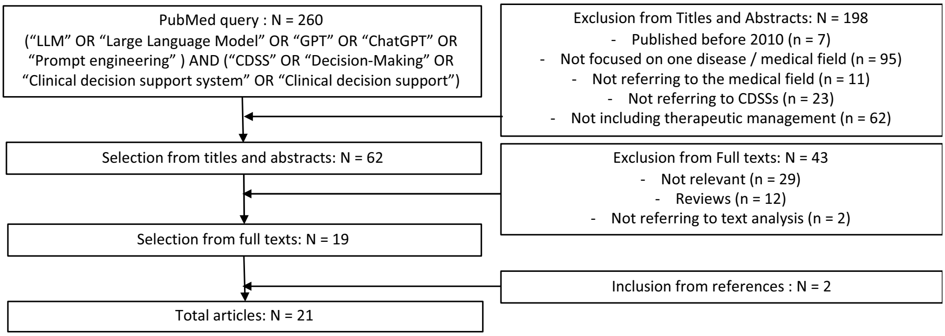
PubMed query : N = 260
("LLM" OR "Large Language Model" OR "GPT" OR "ChatGPT" OR "Prompt engineering") AND ("CDSS" OR "Decision-Making" OR "Clinical decision support system" OR "Clinical decision support")
Exclusion from Titles and Abstracts: N = 198
- Published before 2010 (n = 7)
- Not focused on one disease / medical field (n = 95)
- Not referring to the medical field (n = 11)
- Not referring to CDSSs (n = 23)
- Not including therapeutic management (n = 62)
Selection from titles and abstracts: N = 62
Exclusion from Full texts: N = 43
- Not relevant (n = 29)
- Reviews (n = 12)
- Not referring to text analysis (n = 2)
Selection from full texts: N = 19
Inclusion from references : N = 2
Total articles: N = 21

Figure 1. PRISMA illustrating the selection of relevant articles

4. Discussion

Our review studied the use of LLMs to evaluate their performance as CDSSs. We found that most studies utilized LLMs to respond to prompts related to clinical cases (12/21). Using clinical cases is indeed crucial for assessing LLMs' accuracy. However, effectiveness of LLMs' responses may vary based on the phrasing of the prompts [2] and the origin of clinical cases. Moreover, studies involving real patients (5/12) exhibited lower performance (ranging from 16% to 83%) as compared to results observed with hypothetical patients (58% to 98%), primarily due to patient data privacy concerns. To enhance data safety and reliability, adopting open-source models like LLama is indeed recommended [3], yet only three studies did so, #2, #4 and #20. This may indicate that broader adoption of open-source models requires further enhancement.

Selected studies span various medical specialties, predominantly in oncology (9/21), where ChatGPT's performance, with a precision rate ranging from 16.05% (study #9) to 91.7% (study #17), surpasses the one observed in neurology, with a precision rate ranging from 8% (study #21) to 59.8% (study #15). Nonetheless, these results remain lower than those observed in other medical specialties such as rheumatology or urology, where LLMs' compliance rates range from 83% to 91.6%.

One-third of oncology studies focus on breast cancer (#13, #9, and #5), highlighting contrasting results on the effectiveness of ChatGPT. Study #13 shows a 70% compliance rate of ChatGPT on real cases of invasive ductal carcinoma. Study #9 reported a significantly lower compliance rate of 16.05% in real cases of early-stage breast cancer, and study #5 had a 58.8% compliance rate using fictional patients. These findings underscore that, as compared to guideline-based CDSSs (GL-CDSSs) like Oncodoc2 [4] or DESIREE [5], LLMs have not yet attained their full effectiveness (compliance rate of 91.7% for Oncodoc2 computed on a set of 1624 real patients [4]). Additionally, LLMs are poorly explainable and their validity is questionable as compared with GL-CDSSs.

The overall perception of the use of ChatGPT as a CDSS reveals mixed opinions: seven studies are favorable, six adopt a tempered position, four are unfavorable, and four remain neutral. The cross-analysis (as shown in Table 2) reveals that, although performances are quite varied (seven papers have a high performance and six have a low performance), perceptions remain generally favorable or tempered. Despite this, even studies with medium or low measured performance have recognized the potential of LLMs as CDSSs, one being explicitly favorable despite poor results (#15), and three without any results (#3, #8, #11).

Table 1. Synthesis matrix of articles studied.

#	LLMs	Medical specialty	Gold standard	Prompt for LLMs	Patient type [nb]	Result	PMC/PM Pub Date
				Full text			
1	GPT3.5	Oncology	Not specified	Clinical cases	Fictitious [1]	🟥 -	PMC10544698 (09/2023)
2	5LLMs*	Oncology	Tumor board	Clinical cases	Fictitious [10]	🟨 < 25.0%	PMC10656647 (11/2023)
3	GPT3.5	Ophthalmology	Not specified	General	Fictitious [1]	🟩 -	PMC10362525 (06/2023)
4	7 LLMs**	Neurology	Guidelines	Clinical cases	Fictitious [1]	🟨 -	PMC10840049 (02/2024)
5	GPT3.5	Oncology	Tumor board	Clinical cases	Fictitious [5]	🟥 58.8%	PMC10608120 (10/2023)
6	GPT3.5	Oncology	Tumor board	Clinical cases	Real [20]	🟩 80.0%	PMC10314415 (06/2023)
7	GPT4	Oncology	Clinicians	Guidelines	Fictitious	🟨 88.9%	PMC10722294 (12/2023)
8	GPT-3.5/4	Psychiatry	Guidelines	Clinical cases	Fictitious [8]	🟩 -	PMC10582915 (10/2023)
9	GPT3.5	Oncology	Tumor board	Clinical cases	Real [10]	🟨 16.05%	PMC10579162 (07/2023)
10	GPT4	Rheumatology	Clinicians	Clinical cases	Real [100]	⬜ 83.0%	PMC10684473 (11/2023)
11	GPT3.5	Orthopedics	Guidelines	Guidelines	No patients	🟩 -	PMID:37560946 (08/2023)
12	2 LLMs***	Gastroenterology	Guidelines	Patient	No patients	⬜ 87.0%	PMC10847895 (01/2024)
13	GPT-3.5	Oncology	Tumor board	Clinical cases	Real [10]	🟨 70.0%	PMC10229606 (05/2023)
14	GPT-3.5	Oncology	Guidelines	Clinical cases	Fictitious [1]	⬜ 77.0%	PMC10200252 (04/2023)
15	GPT-4	Neurology	Clinicians	Clinical cases	Real [102]	🟩 59.8%	PMID:38184368 (01/2024)
16	GPT-3.5/4	Urology	Clinicians	Guidelines	Fictitious [25]	🟩 91.6%	PMID:37722842 (09/2023)
17	GPT3.5	Oncology	Guidelines	Guidelines	Fictitious [68]	🟩 91.7%	PMID:38421392 (02/2024)
18	GPT3.5	ORL	Clinicians	Clinical cases	Fictitious [20]	🟨 80.0%-98.0%	PMID:38345613 (02/2024)
19	GPT3.5	Urology	Not specified	General	No patients	🟥 -	PMID:38386789 (01/2024)
20	7 LLMs****	Endocrinology	Guidelines	Guidelines	No patients	⬜ 7.6%$^{GPT3.5}$ 31.0%^{GPT4}	PMID:38419470 (02/2024)
				Abstract only			
21	GPT-3.5	Neurology	Guidelines	Guidelines	No patients	🟥 8.0%	PMID:38124357 (12/2023)

*GPT3.5, GPT4, Galactica, Perplexity, BioMedLM, / ** Bard, PaLM, Bing, GTP3.5, GTP4, Llama, Claude-2, *** GPT4, Google Bard / **** GPT3.5, GPT4, GPT4 Turbo, Google Bard, Bing AI, Perplexity, Claude-2.

The analysis of publication dates indicates that all included studies were published between 04/2023 and 02/2024, shortly after the publicized launching of ChatGPT in late 2022. This suggests that favorable perceptions might be explained by novelty rather than by thorough evaluations. ChatGPT's efficiency has exhibited enhancements, with a performance improvement of 18% for GPT3.5 within a single month, as reported in study #18. Additionally, GPT4 has shown superiority over GTP3.5, with a precision rate increase from 7.6% to 31% for the same task (study #20), thereby underscoring the continuous progress of OpenAI's models. Moreover, in September and October 2023, two separate studies (#8 and #16) initiated comparisons between ChatGPT versions 3.5 and 4. Subsequently, studies #2, #4, #12 and #20 broadened their analytical scope to a wider variety of LLMs and open-source models at large. Thus, we recommend standardizing the evaluation of LLMs, particularly by establishing a common gold standard, for a more comprehensive assessment.

Table 2. Cross-evaluation of ChatGPT's performance and healthcare professionals' perceptions

Perceptions / Performance	No	Neutral	Temper	Yes	Total
Low	2	1	2	1	6
Medium	-	1	1	-	2
High	-	2	2	3	7
No result	2	-	1	3	6
Total	4	4	6	7	21

5. Conclusions

In conclusion, our review reveals there is a true potential of LLMs like ChatGPT as CDSSs, despite challenges in integration and variable effectiveness according to different medical contexts. The improving performance of ChatGPT, particularly in oncology, suggests a positive future role for LLMs in this medical specialty. However, the successful adoption of LLMs in clinical practice will depend on their capacity to addressing reliability, explainability, and ethical concerns. Embracing both CDSS and LLM approaches should not only mitigate these concerns but also enhance patient care quality by combining precision of AI with the nuanced understanding of physicians.

References

[1] Novikava N, Redjdal A, Bouaud J, Séroussi B. Clinical Decision Support Systems Applied to the Management of Breast Cancer Patients: A Scoping Review. Stud Health Technol Inform. 2023 Jun 29;305:353-356. doi: 10.3233/SHTI230503

[2] Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res. 2023;25:e50638. doi: 10.2196/50638.

[3] Toma A, Senkaiahliyan S, Lawler PR, Rubin B, Wang B. Generative AI could revolutionize health care - but not if control is ceded to big tech. Nature. 2023;624:36–38. doi: 10.1038/d41586-023-03803-y

[4] Séroussi B, Laouénan C, Gligorov J, Uzan S, Mentré F, Bouaud J. Which breast cancer decisions remain non-compliant with guidelines despite the use of computerised decision support? British Journal of Cancer. 2013;109:1147. doi: 10.1038/bjc.2013.453

[5] Bouaud J, Pelayo S, Lamy J-B, Prebet C, Ngo C, Teixeira L, Guézennec G, Séroussi B. Implementation of an ontological reasoning to support the guideline-based management of primary breast cancer patients in the DESIREE project. Artif Intell Med. 2020;108:101922. doi: 10.1016/j.artmed.2020.101922 [1]