Can large language models' clinical decision-making match human consensus on when to perform a kidney biopsy?

Michael Toal, Christopher Hill, Michael Quinn, Ciaran O'Neill, Alexander Peter Maxwell

Submitted to: Journal of Medical Internet Research
on: March 07, 2025

Table of Contents

Can large language models' clinical decision-making match human consensus on when to perform a kidney biopsy?

Michael Toal[1, 2] MBChB; Christopher Hill[3] MD; Michael Quinn[2] MD; Ciaran O'Neill[2] PhD; Alexander Peter Maxwell[2] MD, PhD

[1] Queen's University Belfast Belfast GB
[2]Grosvenor Road Royal Victoria Hospital Queen's University Belfast Belfast GB
[3] Lisburn Road Belfast City Hospital Belfast GB

Corresponding Author:
Michael Toal MBChB

Queen's University Belfast
Royal Victoria Hospital
Grosvenor Road
Belfast
GB

Abstract

Background: Artificial intelligence (AI) and Large Language models (LLMs) are increasing in sophistication and have become integrated into many industries. The potential for LLMs to augment clinical decisions is an evolving area of research.

Objective: This study compared the responses of over 1000 kidney specialist physicians (nephrologists) to outputs of commonly used LLMs using a questionnaire determining when a kidney biopsy should be performed.

Methods: This research group completed a large online questionnaire for nephrologists to determine when a kidney biopsy should be performed. The same questions were put to both human doctors and LLMs in an identical order. Eight LLMs were interrogated: Chat GPT 3.5, Mistral Hugging Face, Perplexity, Microsoft Co-pilot, Llama 2, GPT 4.0, MedLM and Claude 3. The most common response given by clinicians (human mode) to each question was taken as the baseline for comparison. Questionnaire responses generated a score reflecting biopsy propensity.

Results: Chat GPT 3.5 and GPT 4.0 had the highest levels of agreement, with the human mode selected in 6/11 questions and a similar propensity score to the human mode. Llama 2 and Microsoft Co-pilot produced similar propensity scores, but with lower levels of agreement with the human mode.

Conclusions: LLM outputs were able to replicate human clinical decision making in this study, however the performance varied widely between LLM models. Questions with more uniform human responses produced LLM outputs with greater alignment, whereas in questions with low levels of human consensus there was poor output alignment. This may limit the practical use of LLMs in real world clinical practice.

(JMIR Preprints 07/03/2025:73603)
DOI: https://doi.org/10.2196/preprints.73603

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Can large language models' clinical decision-making match human consensus on when to perform a kidney biopsy?

Authors: Michael P. Toal[1] MB BCh, Christopher J. Hill[2] MD, Michael P. Quinn[1] MD, Ciaran O'Neill[1] PhD, Alexander P. Maxwell[1] PhD

Institutions:
[1]Centre for Public Health, Grosvenor Road, Queen's University Belfast, Belfast, BT12 6BA Northern Ireland.
[2]Regional Centre for Nephrology and Transplantation, Belfast City Hospital, Belfast, BT9 7AB, Northern Ireland.
Corresponding author: Dr Michael Toal, email:mtoal11@qub.ac.uk

Abstract

Background
Artificial intelligence (AI) and Large Language models (LLMs) are increasing in sophistication and have become integrated into many industries. The potential for LLMs to augment clinical decisions is an evolving area of research.

Objective
This study compared the responses of over 1000 kidney specialist physicians (nephrologists) to outputs of commonly used LLMs using a questionnaire determining when a kidney biopsy should be performed.

Methods
This research group completed a large online questionnaire for nephrologists to determine when a kidney biopsy should be performed. The same questions were put to both human

doctors and LLMs in an identical order. Eight LLMs were interrogated: Chat GPT 3.5, Mistral Hugging Face, Perplexity, Microsoft Co-pilot, Llama 2, GPT 4.0, MedLM and Claude 3. The most common response given by clinicians (human mode) to each question was taken as the baseline for comparison. Questionnaire responses generated a score reflecting biopsy propensity.

Results

Chat GPT 3.5 and GPT 4.0 had the highest levels of agreement, with the human mode selected in 6/11 questions and a similar propensity score to the human mode. Llama 2 and Microsoft Co-pilot produced similar propensity scores, but with lower levels of agreement with the human mode.

Conclusions

LLM outputs were able to replicate human clinical decision making in this study, however the performance varied widely between LLM models. Questions with more uniform human responses produced LLM outputs with greater alignment, whereas in questions with low levels of human consensus there was poor output alignment. This may limit the practical use of LLMs in real world clinical practice.

Keywords

Kidney biopsy, nephrology, large language models, artificial intelligence, decision support

Background

Artificial intelligence in healthcare

The rapid expansion of artificial intelligence (AI) has impacted numerous industries over recent decades. This technology aims to improve efficiency, however there are concerns that human roles may be replaced, and autonomous AI could cause significant disruption to societies.[1,2] AI is now seamlessly integrated into everyday life and common functions such as predictive text messaging, review summaries and customer service chatbots rely on this technology. Generative AI (GAI) can be used to rapidly synthesise images and text, which require academic institutions to consider how to effectively undertake assessments. Large Language Models (LLM) such as Chat GPT, Co-pilot and Llama have rapidly proliferated to employ these developments for personal or professional use.

AI has also been used effectively in healthcare and further expansion is predicted in the years ahead.[3,4] There are many demands on healthcare resources across the world[5] and AI offers opportunities to automate routine human tasks, allowing human practitioners to use their time more effectively on complex problems that are currently beyond the scope of AI.[6,7] In some circumstances, chatbot outputs have even been found to be of higher quality and convey deeper empathy than human responses.[8,9] The diagnostic specialties offer a good template for this illustration. In pathology, AI has been used to rapidly characterise the nature of lesions for rapid detection and coding, allowing the pathologist to analyse specimens with greater efficiency.[10] In radiology, similar pattern recognition has been used to quickly identify abnormalities, as well as helping the radiologist prioritise their workflow, so that the most abnormal or urgent scans are reported first.[7]

Limitations of artificial intelligence

The rise in AI usage does bring significant concerns. Intelligence is not equivalent to

wisdom and AI outputs are reliant on the data used to train these models. Although GAI can produce a detailed response, one criticism is that outputs lack the ‘common sense’ of humans.[11] LLMs can generate false information in the form of hallucinations and produce gender or racially biased outputs.[2,12,13] LLMs are sensitive to phrasing and can generate errors by varying the order of words.[14] How each company trains their LLMs remains confidential, and this lack of transparency is another cause of concern.[3] Using AI within healthcare is dependent on the alignment with human values to establish trust from service users, which is another challenge of any new technology, especially given the issues discussed.[15]

Clinical decision support

Medical practitioners make numerous clinical decisions throughout their working day. How they make these decisions remains poorly understood and open to many potential biases and influences.[16,17] Given that some of these decisions may stand between life and death, harnessing AI to assist physicians in making the best clinical decisions for each patient based on the available body of evidence may represent an opportunity to improve efficiency and enhance safe patient care. LLM outputs have been found to be superior to junior surgical residents’ clinical decision making, but inferior to senior colleagues, however in these studies LLM outputs were limited by inconsistencies and inaccuracies.[18,19] LLMs have been shown to easily pass high stakes written medical examinations such as the USMLE (United States Medical Licensing Examination)[20] and the MRCP(UK) (Membership of the Royal College of Physicians)[21], however appear to perform poorly in questions on rare diseases, perhaps due to a paucity of training data.[22] In the Polish nephrology specialty examination, GPT4 performed at a level similar to the average human candidate, but below that of the top candidates.[23]

Our research group has completed a large international survey of physicians’ clinical decision making, recruiting over 1000 doctors from 83 countries to complete a short online questionnaire.[24] In this study, nephrologists (kidney specialists) were asked to determine when a kidney biopsy was required using clinical vignettes of potential indications and contraindications. A kidney biopsy is used to define which type of kidney disease a patient has so that appropriate treatment can be administered. A biopsy is an invasive investigation with a small but significant risk of serious bleeding complications.[25] The use of AI in nephrology is increasing with recent studies assessing LLM usage for guideline adherence, dialysis management and specialist examinations.[26–29]

We aimed to compare the responses of over 1000 human doctors to LLM outputs, using the same questions on biopsy practice in the same order, to determine if AI can be used as a clinical decision tool in a safe and effective manner.

Methods

Questionnaire design

The detailed methods for the questionnaire have been described in a previous paper.[24] In brief, the questionnaire generated a kidney biopsy propensity score based on the responses to eleven questions, which allowed comparisons between kidney disease specialists to determine if they were more or less likely to recommend this investigation when placed in an identical situation. A higher score was associated with a greater inclination to recommend biopsy in each scenario. For each of these questions,

respondents were asked to select one of five possible responses to each prompt. For the clinical vignettes on indications, this was on a Likert scale from Definitely yes to Definitely no, and for contraindications, by defining a threshold of acceptable risk for a parameter associated with bleeding complications. The most common response (mode) for each question given by human respondents was determined to be the baseline for a comparison with LLM outputs. For each question, the mode was selected by a minimum of 345 and a maximum of 728 human clinicians.

Large language model application

Responses for this human questionnaire were collected from August 2023 to January 2024. LLMs were interrogated from March 2024 to June 2024. At this time, results were not publicly available and therefore could not act as part of an evidence base for the LLM to generate responses. The questions put to the LLM were identical (except for removing the words, 'in your opinion') to those given to human clinicians. They were also presented in the same order.

A propensity score was generated for each LLM using the same parameters as human respondents. Therefore, a LLM which generated a higher score would be more inclined to recommend this investigation and therefore less risk averse. By contrast, a lower score would be indicative of being less inclined to recommend this investigation and more risk averse. A LLM was determined as being a perfect match to human clinicians if the answer selected was identical to the mode in the human questionnaire.

Results

Human respondents' characteristics

A total of 1181 clinicians from 83 countries participated in the study. A summary of clinician characteristics is given in Table 1. The study was open to nephrology trainees and fellows who comprised 14.3% of the total cohort.

	Frequency	%
Sex		
Male	753	64.3
Female	408	34.8
Prefer not to say	9	0.8
Non-binary/Third gender	2	0.2
Age		
20-29	30	2.5
30-39	442	37.5
40-49	327	27.7
50-59	251	21.3
60 or over	130	11.0
Current job title		
Trainee/Fellow	168	14.3
Associate Specialist/Specialty Doctor	122	10.4
Consultant / Attending Physician	733	62.2
Clinical Director or Professor	154	13.1
Other	1	0.1
Continent of Practice		
Europe	405	34.4
North America	352	29.9
South America	85	7.2
Asia	216	18.3

Africa	67	5.7
Oceana	54	4.6

Table 1. Characteristics of human participants

The United States was the largest single national group and 43 states were represented in this cohort. The four devolved nations in the UK were also represented in the second largest cumulative group. 13 nations had greater than 20 clinicians included.

LLM interrogation

A total of eight LLMs were interrogated as detailed in Table 2. The full transcripts of dialogue are detailed in the supplement. The outputs produced by the LLMs varied greatly in terms of detail, however each of the LLMs were instructed to choose an answer from five options. Some LLMs selected more than one answer to certain prompts, refused to give an answer or produced incomplete sentences, however in most instances the question was answered as instructed. An introductory prompt for context was added for GPT 4.0. The programs that were free to use without subscription were interrogated by the first author. GPT 4.0 and Med LM were not freely available, therefore an additional operator with access was employed to reproduce these methods.

LLM	**Date of interrogation**	**Availability**
Open AI: Chat GPT 3.5	27th March 2024	Free without subscription
Mistral Hugging Face	28th March 2024	Free without subscription
Perplexity	28th March 2024	Free without subscription
Microsoft Co-pilot	28th March 2024	Free without subscription
Llama 2 13b chatbot	3rd April 2024	Free without subscription
Open AI: GPT 4.0	22nd April 2024	Subscription
Med LM	26th April 2024	Subscription
Claude 3	13th June 2024	Free without subscription

Table 2. LLMs used and dates of interrogation

LLM prompts

Clinicians were asked if, in their opinion, a kidney biopsy was required in the setting of seven fictional clinical vignettes. All cases were adults with unexplained abnormalities in kidney function, reported as estimated glomerular filtration rate (eGFR) and/or urinary tests (haematuria or proteinuria quantified as grams per day). Four cases were a first presentation to a nephrologist and in three there was a dynamic change over the course of a year.

The determination of when clinicians felt the risk of kidney biopsy outweighed the benefits was explored in a section on potential contraindications, particularly relating to bleeding risk. In the first section, clinicians were presented with five options and asked for the limits of acceptable parameters to proceed to biopsy. This could be the minimum level (e.g. haemoglobin) or maximum level (e.g. systolic blood pressure). The question prompts given to LLMs are detailed in Table 3.

Question code	**Full question**
Q1	Is a renal biopsy required for an adult in the first detection of an unexplained nephrotic

	syndrome of proteinuria 4g/day, peripheral oedema and eGFR>60 ml/min/1.73m^2? Choose from Definitely yes, Probably yes, Unsure, Probably not, Definitely not
Q2	Is a renal biopsy required for an adult in the first detection of unexplained non-visible haematuria, 2g/day of proteinuria and eGFR 40? Choose from Definitely yes, Probably yes, Unsure, Probably not, Definitely not
Q3	Is a renal biopsy required for an adult in the first detection of unexplained non-visible haematuria, 2g/day of proteinuria and eGFR 20 with normal kidney appearances on ultrasound? Choose from Definitely yes, Probably yes, Unsure, Probably not, Definitely not
Q4	Is a renal biopsy required for an adult in the first detection of unexplained non-visible haematuria, 2g/day of proteinuria and eGFR 20 with reduced kidney size on ultrasound? Choose from Definitely yes, Probably yes, Unsure, Probably not, Definitely not
Q5	Is a renal biopsy required for an adult with an unexplained rise in proteinuria from 0.5 to 2g/day in one year with an eGFR > 60? Choose from Definitely yes, Probably yes, Unsure, Probably not, Definitely not
Q6	Is a renal biopsy required for an adult with an unexplained fall in eGFR from 55 to 40 in one year with proteinuria stable at 0.5g/day? Choose from Definitely yes, Probably yes, Unsure, Probably not, Definitely not
Q7	Is a renal biopsy required for an adult with an unexplained fall in eGFR from 55 to 40 AND rise in proteinuria from 0.5 to 2 g/day in one year? Choose from Definitely yes, Probably yes, Unsure, Probably not, Definitely not
Q8	What is the minimum acceptable Haemoglobin for native renal biopsy? Choose from 100g/l, 90 g/l, 80 g/l, Other (please specify) and No minimum level
Q9	What is the minimum acceptable Platelet count for native renal biopsy? Choose from 150 x109, 100 x109, 50 x109, Other (please specify), No minimum level
Q10	What is the maximum acceptable International Normalised Ratio for native renal biopsy? Choose from 1.2, 1.4, 1.6, Other (please specify) and No maximum level
Q11	What is the maximum acceptable Systolic Blood Pressure for native renal biopsy? Choose from 140 mmHg, 160 mmHg, 180 mmHg, Other (please specify) and No maximum level

Table 3. Question prompts given to LLMs.

Comparing human doctor and LLM responses

In the human questionnaire, all available options were selected by at least 3 and at most 728 human doctors. Therefore, none of the LLM outputs could be considered beyond the reach of what a human doctor may select. The subject of kidney biopsy was chosen for this clinician questionnaire as it was subjective, therefore there are a range of acceptable answers. The mode of each answer represents varying proportions of responses to each question. For question 8 (Q8) there was no clear consensus and the mode was selected by only 31.6% of respondents, however for question 9 (Q9) the consensus was clearer and 66.5% of respondents selected the mode. Responses are detailed in Table 4.

The range of agreement between the mode of human responses and LLM outputs ranged from 0 out of 11 (Mistral Hugging Face) to 6 out of 11 (Chat GPT 3.5 and GPT 4.0). 4 of 8 LLMs generated a biopsy propensity score that was equal to or within one point of the mode of human responses.

Using this propensity score as a surrogate marker for clinical risk aversion, the most risk averse LLM output was MedLM, which produced outputs equivalent to the lowest 1% of biopsy propensity scores in human respondents. By contrast, the Claude 3 output produced the highest biopsy propensity score indicating the lowest level of risk aversion, with a score higher than 99% of human respondents. For both MedLM and Claude 3, there was a reasonable agreement between outputs and human responses with 4 or 5 perfect matches out of 11, however the overall approach to risk as indicated by the propensity score was not typical of human responses.

In terms of which LLM most accurately represented human doctor responses, the two Open

AI LLMs Chat GPT 3.5 and GPT 4.0 were the optimal programs for agreement with the human mode and profile of risk aversion, as indicated by the propensity score.

uestion	Humans	Chat GPT 3.5	Mistral Hugging Face	Perplexity	MS Co pilot	Llama 2 13b chat	GPT 4.0	MedLM	Claude 3
1	DY (58.1%)	PY	NA	DY	PY	U	DY & PY	PY	DY
2	DY (59%)	DY	NA	DY	PY	PY	PY	PN	PY
3	DY (51.2%)	DY	NA	DY	PY	U	DY	PN	DY
4	PN (51.4%)	DY	NA	DY	PY	DY	PN & DN	PN	DY
5	PY (47.0%)	PN	NA	PY	U	PY	PY	PN	PY
6	PN (36.1%)	PN	NA	PY	PY	U	U	PN	PY
7	PY (51.7%)	PY	NA	DY	PY	U & PY	DY	PY	DY
8 b (g/l)	90 (31.6%)	100	80	Other	Other	80	100	90	80
9 at (x10^{9})	100 (66.5%)	100	50	50	100	150	100	100	50
10 R	1.2 (51.1%)	1.4	1.4	1.5	1.5	1.4	1.4	1.4	1.5
11 BP (mmHg)	160 (54.7%)	160	140	140	140	140	160	140	160
tal score	23	23	N/A	29	23	22	22-24	11	34
greement		6/11	0/11	4/11	2/11	2/11	6/11	5/11	4/11

Table 4. DY- Definitely Yes, PY- Probably Yes. U-Unsure, PN- Probably Not, DN-Definitely Not, NA- No answer given. Hb- haemoglobin, Plat- platelet count, INR- international normalised ratio, SBP- systolic blood pressure. Red text- most common response (mode) given by human participants. Red text in brackets- the proportion of human respondents in who selected the mode for each question. Green text: LLM output contains mode of human responses.

Discussion

Principal results

In our study, the clinical decision making of doctors specialising in nephrology could be accurately replicated by LLMs. The degree of fidelity differed amongst LLMs and the Open AI models Chat GPT 3.5 and GPT 4.0 produced outputs that were most consistent with typical clinician responses. Similar to our study of human clinicians, most of the LLMs interrogated opted for a balanced approach to this dilemma, producing comparable responses of when to perform a biopsy and when to avoid this procedure.

There were varying degrees of agreement amongst human respondents, with the mode selected by between 31.6% to 66.5% of humans. This consensus was fairly well replicated amongst LLMs. In the question where only 31.6% of human respondents selected the mode, the mode was selected by only 1 out of 8 LLMs, the lowest in our study. Conversely,

in the question with the highest agreement, where 66.5% of humans selected the mode, this human mode was also selected by 4 out of 8 LLMs, the joint highest in our study. This suggests that the grey areas of ambiguity in clinical decision making can also be reflected in the LLM outputs. One potential use for this technology would be to assist the clinician resolve an uncertain decision, however in this instance, this uncertainty is also reflected in LLM outputs, limiting utility in real-world practice.

The propensity to perform an invasive procedure is inevitably linked to the tolerance of potential risks. Therefore, we used this score as a surrogate marker for risk aversion in human clinicians. When this is applied to LLMs there is variable risk aversion in these models. MedLM outputs were the most risk averse, as there was a higher threshold to perform a kidney biopsy, as well as a low tolerance for potential contraindications that would increase the risk of a bleeding complication. In contrast, the outputs for Claude 3 were the least risk averse, as every clinical vignette was met with a response that a kidney biopsy was definitely or probably required and the loose limits for potential contraindications could be considered by some clinicians to be reckless.

Limitations

This study has limitations which should be addressed. Firstly, this is a small sample of eleven questions used to interrogate LLMs, therefore caution is required in interpretation. Secondly, this study assessed the LLMs' ability to make decisions based on short case vignettes and this may not necessarily be generalisable to more complex "real-life" clinical scenarios, as LLM accuracy has been shown to be poorer on longer questions.[21]

Strengths

This study also has notable strengths. Human decision-making is poorly understood and clinical decisions should be based on integrating the best available evidence for the care of an individual patient. AI-assisted decision aids are rapidly expanding into medicine and this is the first study that the authors are aware of which compares a large number of human respondents to LLM outputs when placed in identical scenarios.

Implications for future research

There has been a rapid proliferation of medical research into the uses of AI in healthcare and how these are best integrated into clinical practice remains unclear. As LLMs continue their journey of increased sophistication and accuracy, AI assistance will likely become integral to all aspects of life. How we best apply this in healthcare remains a challenge to be addressed in upcoming years, however it is important that LLM outputs align with human values, which can be shaped by supervised reinforcement learning by expert physicians and patients.[15]

Conclusions

Large language models can accurately replicate human clinical decision making when short questions are inputted. There is variable performance in these models, however Chat GPT 3.5 and GPT 4.0 outputs were the most consistent with humans in our study. Caution should be applied when considering how these can be used to assist clinicians, as there remains many unanswered questions as to how physicians should use these tools for safe

and effective patient care.

Acknowledgments
The authors would like to thank Mr Marc McNicholl and Mr Tushar Gandhi for their assistance in reproducing the applied methods for paid services.

Funding Information
MT is supported by a clinical research fellowship award from the Northern Ireland Kidney Research Fund (NIKRF). This organisation had no input into the design or conduct of this study.

Conflicts of Interest
The authors report no relevant conflicts of interest.

Abbreviations
AI: Artificial Intelligence
DN: Definitely no
DY: Definitely yes
eGFR: Estimated glomerular filtration rate
g: grams
GAI: Generative artificial intelligence
GPT: Generative pre-training transformer
Hb: Haemoglobin
INR: International normalised ratio
l: litre
LLM: Large Language Model
m: metre
mmHg: millimetres of mercury
min:minute
ml: millilitres
MRCP: Membership of the Royal College of Physicians
MS: Microsoft
NA: No answer given
Plat: Platelet count
PN: Probably no
PY: Probably yes
Q: Question
SBP: Systolic blood pressure
U: Unsure
USMLE: United States Medical Licencing Examination

Data availability
Data is available upon reasonable request by contacting the corresponding author.

References

1. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. J R Soc Med. 2023 May 1;116(5):181–2.
2. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024 Sep 30;9:2613–22.
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Aug 1;29(8):1930–40.
4. Thirunavukarasu AJ. How Can the Clinical Aptitude of AI Assistants Be Assayed? J Med Internet Res. 2023 Dec 5;25(e51603).
5. Kohane IS. Large Language Models and the Degradation of the Medical Record. N Engl J Med [Internet]. 2024 Oct 31;391(17):1564–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/39465898
6. Urquhart E, Ryan J, Hartigan S, Nita C, Hanley C, Moran P, et al. A pilot feasibility study comparing large language models in extracting key information from ICU patient text records from an Irish population. Intensive Care Med Exp. 2024 Dec 1;12(1).
7. Najjar R. Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. Diagnostics (Basel). 2023 Aug 25;13(17):2760.
8. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023 Jun 5;183(6):589–96.
9. Perlis RH, Goldberg JF, Ostacher MJ, Schneck CD. Clinical decision support for bipolar depression using large language models. Neuropsychopharmacology. 2024 Aug 1;49(9):1412–6.
10. Hermsen M, Bel T, Boer M Den, Steenbergen EJ, Kers J, Florquin S, et al. Deep learning-based histopathologic assessment of kidney tissue. J Am Soc Nephrol. 2019;30(10):1968–79.
11. Kejriwal M, Santos H, Mulvehill AM, Shen K, Mcguinness DL, Lieberman H. To find out how smart AI is, first test its common sense. Nature. 2022 Oct 10;634.
12. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health. 2024 Jan 1;6(1):e12–22.
13. Fang X, Che S, Mao M, Zhang H, Zhao M, Zhao X. Bias of AI-generated content: an examination of news produced by large language models. Sci Rep. 2024 Dec 1;14(5224).
14. Salihu A, Gadiri MA, Skalidis I, Meier D, Auberson D, Fournier A, et al. Towards AI-assisted cardiology: a reflection on the performance and limitations of using large language models in clinical decision-making. EuroIntervention. 2023 Dec 1;19(10):e798–801.
15. Yu KH, Healey E, Leong TY, Kohane IS, Manrai AK. Medical Artificial Intelligence and Human Values. N Engl J Med. 2024 May 30;390(20):1895–904.
16. Hall KH. Reviewing intuitive decision-making and uncertainty: The implications for medical education. Med Educ. 2002;36(3):216–24.
17. Sacks GD, Dawes AJ, Tsugawa Y, Brook RH, Russell MM, Ko CY, et al. The Association Between Risk Aversion of Surgeons and Their Clinical Decision-Making. J Surg Res. 2021 Dec 1;268:232–43.
18. Palenzuela DL, Mullen JT, Phitayakorn R. AI Versus MD: Evaluating the surgical decision-making accuracy of ChatGPT-4. Int Surg J. 2024 Aug 1;176(2):241–5.
19. Huo B, Calabrese E, Sylla P, Kumar S, Ignacio RC, Oviedo R, et al. The performance of artificial intelligence large language model-linked chatbots in surgical decision-making for gastroesophageal reflux disease. Surg Endosc. 2024 May 1;38(5):2320–30.
20. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. ArXiv [Internet]. 2023 Mar 20;2303(133752v2). Available from:

http://arxiv.org/abs/2303.13375
21. Maitland A, Fowkes R, Maitland S. Can ChatGPT pass the MRCP (UK) written examinations? Analysis of performance and errors using a clinical decision-reasoning framework. BMJ Open. 2024 Mar 15;14(3).
22. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat Commun. 2024 Mar 6;15(1).
23. Nicikowski J, Szczepański M, Miedziaszczyk M, Kudliński B. The potential of ChatGPT in medicine: an example analysis of nephrology specialty exams in Poland. Clin Kidney J. 2024 Aug 1;17(8).
24. Toal M, Hill C, Quinn M, McQuarrie E, O'Neill C, Maxwell AP. An International Study of Variation in Attitudes to Kidney Biopsy Practice. Clin J Am Soc Nephrol. 2024 Dec 20;19(12).
25. Hogan JJ, Mocanu M, Berns JS. The native kidney biopsy: Update and evidence for best practice. Clin J Am Soc Nephrol. 2016 Feb 5;11(2):354–6.
26. Miao J, Thongprayoon C, Suppadungsuk S, Garcia Valencia OA, Cheungpasitporn W. Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications. Medicina (Lithuania). 2024 Mar 8;60(3).
27. Miao J, Thongprayoon C, Cheungpasitporn W. Assessing the Accuracy of ChatGPT on Core Questions in Glomerular Disease. Kidney Int Rep. 2023 Aug 1;8(8):1657–9.
28. Maursetter L. Will ChatGPT Be the Next Nephrologist? Clin J Am Soc Nephrol. 2023 Dec 1;19(1):2–4.
29. Kotanko P, Zhang H, Wang Y. Artificial Intelligence and Machine Learning in Dialysis Ready for Prime Time? Clin J Am Soc Nephrol. 2023 Jun 1;18(6):803–5.

Supplementary Files

Related publication(s) - for reviewers eyes onlies

Full international human study paper.
URL: http://asset.jmir.pub/assets/361c4fdbee340a7bf3bf9e34155c87a9.pdf