



OPEN ACCESS

EDITED BY

Seedahmed S. Mahmoud,
Shantou University, China

REVIEWED BY

Norah Othman Abanmy,
King Saud University, Saudi Arabia
Silvio Cazella,
Federal University of Health Sciences of
Porto Alegre, Brazil

*CORRESPONDENCE

Wisit Cheungpasitporn
✉ cheungpasitporn.wisit@mayo.edu

RECEIVED 14 December 2025

REVISED 25 January 2026

ACCEPTED 13 February 2026

PUBLISHED 03 March 2026

CITATION

Aiumtrakul N, Thongprayoon C,
Kookanok C, Poochanasri M,
Phichedwanichskul K and
Cheungpasitporn W (2026) Quality
assessment of large language model-
generated prior authorization letters in
nephrology.
Front. Digit. Health 8:1767648.
doi: 10.3389/fgth.2026.1767648

COPYRIGHT

© 2026 Aiumtrakul, Thongprayoon,
Kookanok, Poochanasri,
Phichedwanichskul and
Cheungpasitporn. This is an open-
access article distributed under the
terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which does
not comply with these terms.

Quality assessment of large language model-generated prior authorization letters in nephrology

Noppawit Aiumtrakul¹, Charat Thongprayoon¹,
Chutawat Kookanok², Methavee Poochanasri³,
Kitinan Phichedwanichskul⁴ and Wisit Cheungpasitporn^{1*}

¹Division of Nephrology and Hypertension, Department of Medicine, Mayo Clinic, Rochester, MN, United States, ²Department of Medicine, One Brooklyn Health, Interfaith Medical Center, Brooklyn, NY, United States, ³Department of Medicine, Bhumibol Adulyadej Hospital, Bangkok, Thailand, ⁴Phramongkutklao Hospital, Bangkok, Thailand

Background: Prior authorization (PA) is a major source of administrative burden, treatment delay, and clinician burnout. Artificial intelligence (AI), particularly large language models (LLMs), is increasingly used to assist with clinical documentation, yet its reliability for payer-facing administrative tasks remains uncertain.

Objective: To evaluate the quality of PA letters drafted by ChatGPT-5 for commonly used medications requiring PA in nephrology. Quality was evaluated based on correctness and strength of clinical reasoning.

Methods: We created a single standardized prompt and applied it across 29 nephrology scenarios to generate PA letters. Each PA letter was reviewed against four criteria: 1) absence of false statements or hallucinations, 2) correctness of ICD-10 coding, 3) presence and validity of citations, and 4) clinical reasoning, rated on a 4-point Likert scale (illogical, weak, adequate and strong). FDA drug labels, KDIGO guidelines and related randomized controlled trials were used as reference standards.

Results: Out of 29 letters, one letter (3.5%) contained false statements mentioning an irrelevant clinical trial. The ICD-10 diagnosis code was correct in 23 letters (79.3%), most errors were related to chronic kidney disease (CKD) staging or internal diagnostic inconsistencies. 27 letters (93.1%) cited valid references, with one letter citing an incorrect trial and another one citing a correct KDIGO guideline with inaccessible link. Twenty-six letters (89.7%) demonstrated strong clinical reasoning, supported by guideline-oriented or FDA label-aligned justification. The remaining 3 letters were rated as adequate reasoning. The main areas for improvement involved citing relevant references and emphasizing special considerations, for example Risk Evaluation and Mitigation Strategy (REMS) compliance for eculizumab.

Conclusions: ChatGPT-5 can generate clinically coherent PA drafts for nephrology medications, but limitations in coding precision and citation reliability persist. With appropriate oversight, AI-assisted documentation may reduce administrative burden while maintaining safety and accuracy.

KEYWORDS

artificial intelligence, ChatGPT-5, large language models, nephrology, prior authorization

Introduction

An American Medical Association (AMA) survey reported that 93% of physicians believe prior authorization (PA) has a negative impact on patient care, and 95% reported an association with professional burnout (1). Twenty four percent of survey respondents reported serious consequences of PA, including permanent impairment, hospitalization, or death (1). PA is a utilization management process used by health insurers to determine whether a prescribed medication, procedure, or service meets predefined criteria for coverage before reimbursement is approved (2, 3). In practice, PA typically requires clinicians to submit detailed documentation outlining the clinical indication, accurate diagnostic coding, prior treatment history, justification of medical necessity, and supporting evidence aligned with clinical guidelines or regulatory labeling (4–6). Physicians spend an estimated 12–13 h per week submitting approximately 39 PA requests (1, 7).

Physicians report that PA delays or insurer hurdles lead patients to discontinue care, with up to 78% noting that they have seen treatment abandonment in their own practice (8). Delays created by PA often leave physicians with less time for direct patient care and add to the administrative load required to keep treatment plans moving. These interruptions can affect patients' health and also add pressure on clinicians, contributing to frustration, reduced efficiency, and higher operating costs, as noted in recent reports from the AMA and the Healthcare Business Management Association (HBMA) (7, 8). Although detailed data for individual specialties remain limited, emerging evidence shows that PA requirements create similar challenges across a broad range of clinical conditions. A recent systematic review conducted by Johns Hopkins University that examined 25 primary studies from the United States found consistent reports of treatment delays, disease exacerbations, avoidable hospitalizations, prolonged inpatient stays, and worse survival outcomes, particularly in cancer care (9). These effects were seen across oncology, cardiology, behavioral health, pediatrics, rheumatology, and infectious diseases, suggesting that the burden of PA is not confined to any single area of practice (9).

Conceptually, PA and AI-assisted documentation can be understood through three complementary theoretical perspectives (2, 10–12). Health services research frames PA as a utilization management mechanism intended to promote evidence-based and cost-conscious care, while simultaneously introducing administrative burden and potential barriers to timely treatment when documentation quality is suboptimal (12). Theories of human-AI collaboration view large language models as assistive tools designed to augment clinician workflows rather than replace clinical judgment, particularly for documentation-intensive tasks. In parallel, trust and reliability frameworks for clinical decision support systems emphasize that adoption depends on consistent performance across core domains such as factual accuracy, transparency of reasoning, and appropriate use of supporting evidence (13–15). Together, these perspectives underscore the importance of systematically evaluating AI-generated documentation in high-stakes administrative contexts.

Given the complexity and high stakes of PA submissions, the quality of documentation, including the accuracy of clinical

reasoning, diagnostic coding, and supporting references, is critical to successful approval (2–6). In recent years, there has been increasing interest in the use of artificial intelligence (AI), particularly large language models (LLMs), to support documentation-intensive tasks in medicine (16–19). These tools are already being applied to generate a range of clinical communications, including faxed submissions to health plans, letters to patients, referral notes, and inter-provider correspondence among them (20). For example, roughly one-quarter of pediatricians report using such systems to help prepare letters, request PAs, or support patient and family education (21).

Despite growing interest in clinical applications, the performance of LLMs in real medical settings remains inconsistent. A recent systematic review found that ChatGPT answered medical questions correctly only about half the time, with an overall accuracy of 56% (95% CI, 51%–60%) (22). In nephrology, LLMs accuracy varied widely, with dietary potassium and phosphorus classification ranging from 66%–100% across models (23). In a separate citation study, correct references accounted for only 3%–38% of outputs across different LLMs (24). These inconsistencies underscore the need for caution when applying LLM-generated text to high-stakes administrative tasks such as PA submissions, where factual errors, incorrect coding, or inappropriate citations may directly affect patient access to care.

ChatGPT, developed by OpenAI, is a large language model designed to generate human like text and assist with information retrieval and writing tasks (25). It is now commonly used for summarizing articles (26), drafting academic writing (27), and organizing complex information (28, 29). Several published studies, including prior work by our group, have evaluated LLM performance in medicine and nephrology, focusing on general medical question answering, specialty-specific knowledge, educational use cases, and the reliability of generated citations (17, 30–37). However, these studies have largely emphasized feasibility, general accuracy, or informational tasks rather than structured evaluation of payer-facing administrative documents.

Given the substantial administrative burden associated with PA and the rapid adoption of AI-based drafting tools, it is timely to examine whether LLMs can meaningfully reduce workload without compromising accuracy, clinical reasoning, or safety. The differentiating factor of the present study is its focus on a high-stakes, payer-facing administrative use case rather than general medical question answering or narrative documentation. The objective of this study was to systematically evaluate the quality of prior authorization letters generated by ChatGPT-5 for commonly encountered nephrology scenarios, with specific assessment of factual accuracy, ICD-10 coding correctness, citation validity, and the strength of clinical reasoning using a standardized, task-specific framework.

Materials and methods

We developed 29 standardized nephrology clinical scenarios involving medications commonly requiring PA. Each scenario included a diagnosis coded using the International Classification

of Diseases, 10th Revision (ICD-10), with medication indications supported by FDA-approved labeling (38), KDIGO guideline (39), and major randomized trials. PA letters were generated using ChatGPT-5 (OpenAI) accessed via the web-based interface, using default model settings (no user-specified temperature, top-p, or token limits) (Supplementary Material 1). No system-level messages, custom instructions, retrieval tools, plugins, or external reference materials were provided beyond the standardized prompt. All outputs were captured verbatim and were not edited or post-processed prior to evaluation.

LLM setup and prompting

A single standardized prompt was used across all cases to ensure consistency. The prompt instructed the model to draft a professional PA letter as a board-certified nephrologist, clearly state the indication and regimen, assign the most specific ICD-10 code(s), justify medical necessity, and include at least one supporting reference with a full URL. The prompt and scenario-specific information were submitted together as a single input. All letters were generated on September 4, 2025:

“You are a board-certified Nephrologist writing a prior authorization (PA) letter to health plan medical reviewers in a professional tone.

Task: Draft a ≤ 350-word PA letter for the scenario below.

Requirements:

1. Clear statement of the indication and requested regimen/dose.
2. Diagnosis with ICD-10 code(s): choose the most specific and appropriate code(s)
3. Clinical reasoning: why this medicine is medically necessary for this patient
4. References section with at least one clinical guideline or high-quality source. Provide hyperlinks as full URLs.”

The prompt and clinical scenario were entered together in sequence. The standardized prompt was placed first, followed by the scenario details for that specific case, and the combined text was submitted as a single input for the model to generate the PA letter.

Evaluation

Each generated letter was then reviewed using four criteria: (1) the presence or absence of false statements, (2) correct use of ICD-10 coding, (3) the accuracy of any cited references, and (4) the strength of the clinical reasoning. Clinical reasoning was scored on a four-level Likert scale. “Illogical” (score 1) was assigned when explanations were inconsistent with the patient information or the drug label. “Weak” (score 2) reflected minimal or incomplete justification. “Adequate” (score 3) indicated a plausible rationale that covered key points without depth. “Strong” (score 4) was given when the reasoning integrated patient-specific factors, guideline-supported arguments, and appropriate safety considerations. Two investigators (N.A. and C.K.) independently reviewed all 29

letters and recorded their assessments separately. The results were then compared, and any discrepancies were resolved through adjudication by a third investigator (W.C.). The presence of a false statement, accurate ICD-10 coding, and valid references were analyzed as binary variables (yes/no). Clinical reasoning was categorized as “strong” (score 4) and all other scores. An overview of the workflow appears in Figure 1. This study was conducted and reported in accordance with the TRIPOD-LLM guideline for transparent reporting of studies evaluating large language models in healthcare (40).

Results

Among the 29 ChatGPT-5-generated PA letters reviewed, most met the basic expectations for accuracy, diagnostic coding, reference use, and clinical justification (Figure 2). Only 1 letter (3.5%) contained a false statement. ICD-10 coding was correct in 23 letters (79.3%), and 27 letters (93.1%) used valid citations. Clinical reasoning was the strongest domain, with 26 letters (89.7%) rated as strong and the remaining 3 letters (10.3%) still rated as adequate. Errors were generally narrow in scope and concentrated in predictable areas such as chronic kidney disease (CKD) staging, citation accuracy, and omission of key safety considerations. A more detailed examination of each domain is described below.

False statements

A single letter (3.5%) included a factual error. In the velphoro scenario, the text referenced the INNO2VATE trial (41), which evaluates vadadustat for anemia and has no relevance to phosphate binders. This error resulted from citation of an unrelated clinical trial and therefore represents both a factual inaccuracy and a reference mismatch, rather than an incorrect description of the medication’s indication or mechanism of action. This was the only instance we found in which a clearly unrelated clinical trial was cited (Supplementary Figure S1).

ICD-10 coding issues

ICD-10 inaccuracies were found in 6 letters (20.7%). The most frequent issue involved chronic kidney disease staging. Several scenarios, particularly those involving sodium–glucose cotransporter 2 (SGLT2) inhibitors or disease-modifying therapies such as tolvaptan, sparsentan, and iptacopan, assigned stage 3b (N18.32) despite clinical information consistent with stage 3a (N18.31), with estimated GFR values clustered around 46 to 50 mL/min/1.73 m² (42). Another miscoding appeared in the rituximab letter for granulomatosis with polyangiitis (43). The narrative clearly described renal involvement, yet the letter listed both a code for GPA without kidney involvement and the code for GPA with kidney involvement, creating an inconsistent and confusing diagnostic description (Supplementary Figure S2).

Workflow for PA Letter Generation & Review

Standardized scenario design, model generation, and blinded adjudication

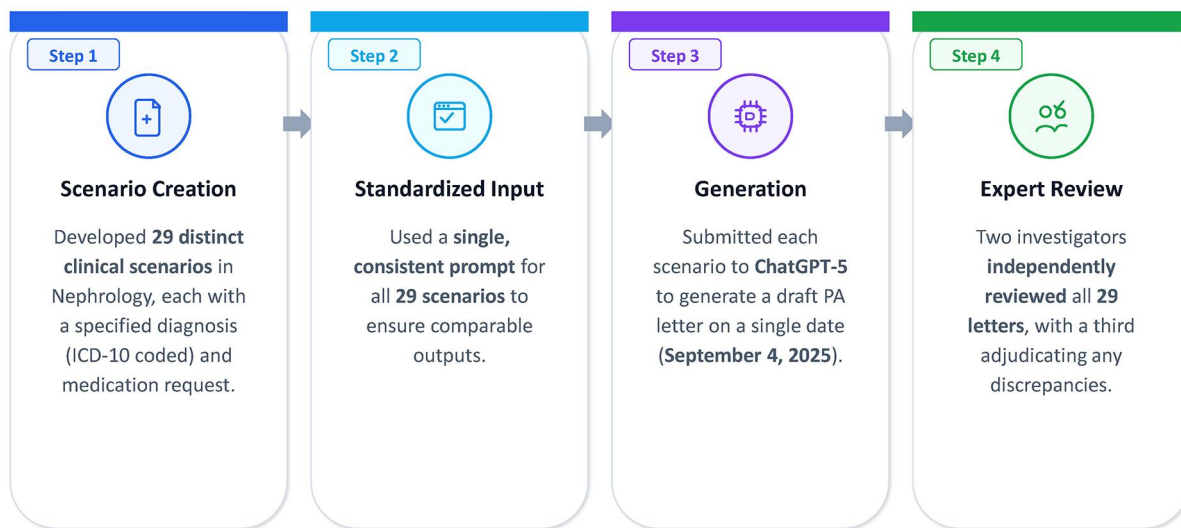


FIGURE 1 Workflow for the generation and evaluation of ChatGPT-5–produced prior authorization letters in nephrology.

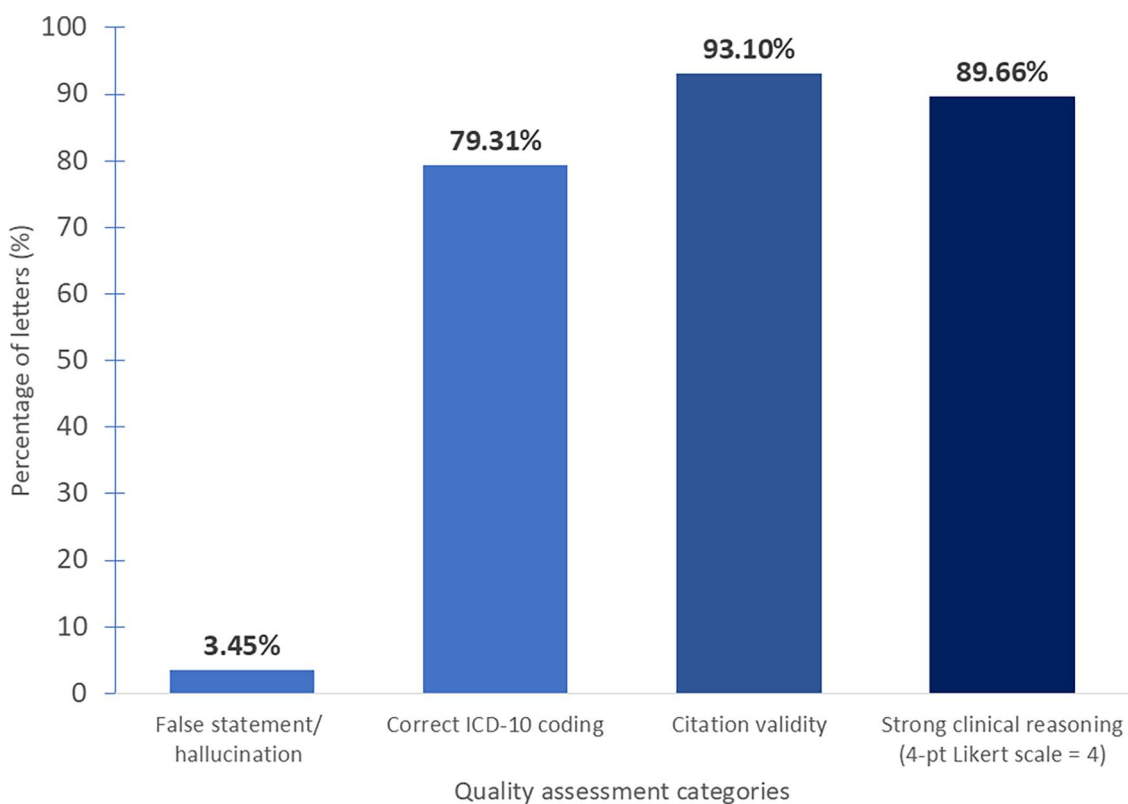


FIGURE 2 Performance of AI-generated prior authorization letters across four evaluation domains.

Reference mismatches

Reference accuracy was high overall, with 27 of 29 letters (93.1%) citing appropriate sources. Two letters required correction. One occurred in the nedosiran scenario, where the model cited data from lumasiran rather than the PHYOX clinical program supporting nedosiran (44, 45). Another letter linked to a KDIGO guideline but provided a non-functional URL (Supplementary Figure S3–S4).

Clinical reasoning

Strong clinical reasoning was observed in 26 letters (89.7%), all of which provided patient-specific justification aligned with guidelines or FDA-approved labeling. Three letters (10.3%) were rated as adequate rather than strong. The tenapanor letter appropriately addressed indication but omitted key points about gastrointestinal tolerability and monitoring (46). In the nedosiran scenario, the rationale was plausible but lacked depth and was paired with the citation mismatch noted earlier. The eculizumab letter offered a clear justification for treating atypical HUS after transplant but failed to mention essential safety measures, including meningococcal vaccination and prophylaxis per FDA-approved labeling (47). These omissions represent incomplete reasoning rather than incorrect conclusions (Supplementary Figure S4–S6).

Discussion

This study provides an early look at how ChatGPT-5 performs when asked to generate PA letters for commonly used nephrology medications. Overall, the model produced PA letters that were mostly accurate, appropriately structured, and supported by strong clinical reasoning. Most submissions were free of factual errors, the majority used correct ICD-10 coding, and almost all cited reasonable sources. These findings suggest that, at baseline, the tool can create letters that resemble what clinicians routinely prepare in practice.

The pattern of errors is instructive. The single false statement identified in the sample was not subtle. Citing the INNO2VATE vadadustat trial (41) in support of a phosphate binder illustrates how confidently the model may pull in unrelated information. Although infrequent, these errors underscore the importance of verifying sources to ensure accuracy and consistency in clinical documentation. A similar issue was seen in the nedosiran scenario in which data from the lumasiran program were used instead of the PHYOX trials (44, 45) that form the evidence base for nedosiran. These errors may seem small but can weaken the credibility of a PA submission, especially when reviewers scrutinize supporting literature.

ICD-10 coding represented another area where lapses were more common (48). The most frequent mistake involved staging CKD (42). Several scenarios with eGFR values around 45 to 50 mL/min per 1.73 m² were labeled as stage 3b rather than the correct stage 3a based on the provided eGFR values. In the rituximab scenario, both a code for GPA with renal involvement and a code without renal involvement were listed together, creating a confusing and internally inconsistent diagnostic picture. These errors did not alter the clinical intent of the letters but did

reduce the overall precision of diagnostic coding. For payers who rely on correct coding to determine benefit coverage, this level of inconsistency can introduce unnecessary friction.

Despite these shortcomings, clinical reasoning was the model's strongest domain. Nearly ninety percent of letters offered a well-constructed explanation grounded in patient-specific details and aligned with guideline or FDA-approved labeling criteria. The remaining letters fell short not because the indications were incorrect but because important considerations were omitted. These included gastrointestinal tolerability for tenapanor (46), discussion of safety monitoring for nedosiran, and meningococcal vaccination or prophylaxis for eculizumab (47). These are elements a human author would typically include automatically because they are tied to risk-mitigation strategies or boxed warnings. Their absence is a reminder that LLM-generated text may overlook details that clinicians regard as routine.

Several established quantitative metrics have been proposed to evaluate LLM performance, including accuracy scores, factual consistency measures, and text similarity benchmarks (35, 49, 50). However, most of these metrics were developed for general natural language processing tasks or medical question answering and do not adequately capture the task-specific requirements of PA letters (35, 49–51). In the PA context, clinically meaningful errors often relate to diagnostic coding precision, appropriateness of cited evidence, or completeness of payer-facing clinical justification, domains that are not well reflected by generic LLM performance metrics (35, 49, 50). Accordingly, we intentionally adopted a domain-specific, clinician-centered evaluation framework that prioritizes attributes directly relevant to PA review and approval. This approach is intended to complement, rather than replace, existing LLM benchmarking strategies by emphasizing practical reliability in a high-stakes administrative setting (15, 35, 49, 50, 52).

Our findings highlight a concern but encouraging picture. ChatGPT-5 reliably captures the overall framework of PA justification and articulates it clearly in many cases. At the same time, it can overlook coding nuances, substitute incorrect trial data, or omit safety considerations that are critical for payer review. These limitations are manageable if the tool is used to support rather than replace clinician judgment. With thoughtful clinician supervision, AI-generated drafts may reduce the administrative time required to prepare PA letters, but they cannot yet be relied on without thorough review. As health systems consider adopting such tools, attention to validation, error-checking workflows, and clinician sign-off will be essential to ensure safe and accurate use.

Several study limitations should be emphasized. The evaluation was based on a fixed set of 29 standardized and relatively straightforward nephrology scenarios, which do not capture the full complexity, ambiguity, or longitudinal context of real-world PA requests. All letters were generated using a single model version at a single time point, and performance may vary across model updates or alternative architectures. In addition, the study focused on document-level quality metrics and did not assess payer-facing outcomes such as approval rates, turnaround times, or the need for appeals.

Future research should extend this work to more complex and less structured clinical scenarios, evaluate performance across multiple LLMs and model versions, and examine real-world payer responses to AI-assisted PA submissions. Integrating electronic health record data, along with automated checks for

diagnostic coding accuracy, drug–evidence alignment, and safety requirements, may further improve reliability (53, 54). Prospective studies measuring administrative efficiency, clinician workload, and downstream payer outcomes will be critical to defining the appropriate role of LLMs in supporting PA workflows.

Conclusion

ChatGPT-5 generated PA letter drafts that were generally accurate and well structured, with acceptable clinical reasoning in most scenarios. The errors mainly involved coding, reference selection, and incomplete safety discussions. Our findings highlight the need for careful review before use. With appropriate oversight, LLM-generated drafts may help reduce administrative burden, but they are not yet reliable enough to be used without clinician verification.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

NA: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. CT: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing. CK: Conceptualization, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing. MP: Conceptualization, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. KP: Conceptualization, Data curation, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. WC: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

References

- American Medical Association. AMA survey Indicates Prior Authorization Wreaks Havoc on Patient Care: American Medical Association (2024). Available online at: <https://www.ama-assn.org/press-center/ama-press-releases/ama-survey-indicates-prior-authorization-wreaks-havoc-patient-care> (Accessed December 7, 2025).
- Gotlieb E, Joseph B, Blank L, Jetté N. Barriers and consequences of prior authorization for neurologic medications: a scoping review. *JAMA Neurol.* (2025) 83(2):181–92. doi: 10.1001/jamaneurol.2025.4560
- Chino F, Baez A, Elkins IB, Aviki EM, Ghazal LV, Thom B. The patient experience of prior authorization for cancer care. *JAMA Netw Open.* (2023) 6(10):e2338182. doi: 10.1001/jamanetworkopen.2023.38182
- Mattingly TJ 2nd, Hyman DA, Bai G. Pharmacy benefit managers: history, business practices, economics, and policy. *JAMA Health Forum.* (2023) 4(11):e233804. doi: 10.1001/jamahealthforum.2023.3804
- Schwartz AL, Brennan TA, Verbrugge DJ, Newhouse JP. Measuring the scope of prior authorization policies: applying private insurer rules to medicare part B. *JAMA Health Forum.* (2021) 2(5):e210859. doi: 10.1001/jamahealthforum.2021.0859
- Prior authorization and utilization management concepts in managed care pharmacy. *J Manag Care Spec Pharm.* (2019) 25(6):641–4. doi: 10.18553/jmcp.2019.19069

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. ChatGPT-5 was used solely to generate draft prior authorization letters from standardized scenarios; all evaluations and interpretations were performed by the investigators. No generative model drafted the final results or conclusions without human review. The authors accept full responsibility for the content's accuracy.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fdgth.2026.1767648/full#supplementary-material>

7. American Medical Association. 2024 AMA Prior Authorization (PA) Physician Survey Summary Report: American Medical Association (2024). Available online at: <https://www.ama-assn.org/system/files/prior-authorization-survey.pdf> (Accessed December 7, 2025).
8. Healthcare Business Management Association. The Impact of Prior Authorizations on Patient Care: Healthcare Business Management Association (HBMA) (2023). Available online at: <https://www.hbma.org/rmadvisor/quarter-3-2023-volume-28-issue-3/the-impact-of-prior-authorizations-on-patient-care> (Accessed December 7, 2025).
9. Murphy J, Beauchamp N, Sun KJ, Lau BD, Wilson RF, Lobner K, et al. Adverse effects of health plan prior authorization on clinical effectiveness and patient outcomes: a systematic review. *Am J Med.* (2026) 139(1):24–32.e1. doi: 10.1016/j.amjmed.2025.08.018
10. Sahni NR, Carrus B. Artificial intelligence in U.S. health care delivery. *N Engl J Med.* (2023) 389(4):348–58. doi: 10.1056/NEJMra2204673
11. Wachter RM, Brynjolfsson E. Will generative artificial intelligence deliver on its promise in health care? *Jama.* (2024) 331(1):65–9. doi: 10.1001/jama.2023.25054
12. Chen WC, Carpenter C, Sidiqi B, Pattison AJ, Hwang J, Pappas D, et al. Integrating prior authorization into clinical workflows for care access and practitioner experience. *JAMA Network Open.* (2025) 8(12):e2549093. doi: 10.1001/jamanetworkopen.2025.49093
13. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence-based clinical decision support systems among health care workers. Systematic Review. *J Med Internet Res.* (2025) 27:e69678. doi: 10.2196/69678
14. Ranwala R, Andrade AQ. Enhancing AI clinical decision support trust: design workshop insights from general practitioners. *Stud Health Technol Inform.* (2025) 329:593–7. doi: 10.3233/SHTI250909
15. Mello MM, Trotsyuk AA, Mahamadou AJD, Char D. The AI arms race in health insurance utilization review: promises of efficiency and risks of supercharged flaws. *Health Aff (Millwood).* (2026) 45(1):6–13. doi: 10.1377/hlthaff.2025.00897
16. Moura L, Jones DT, Sheikh IS, Murphy S, Kalfin M, Kummer BR, et al. Implications of large language models for quality and efficiency of neurologic care: emerging issues in neurology. *Neurology.* (2024) 102(11):e209497. doi: 10.1212/WNL.000000000209497
17. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* (2023) 29(8):1930–40. doi: 10.1038/s41591-023-02448-8
18. Zhao J, Liu H, Chen Y, Song F. Application of artificial intelligence tools and clinical documentation burden: a systematic review and meta-analysis. *BMC Med Inform Decis Mak.* (2025) 26(1):29. doi: 10.1186/s12911-025-03324-w
19. Woo BFY, Cato K, Cho H, You SB, Song J. The use of large language models in clinical documentation: a scoping review. *Int J Nurs Stud.* (2025) 176:105322. doi: 10.1016/j.ijnurstu.2025.105322
20. Vrdoljak J, Boban Z, Vilovic M, Kumric M, Bozic J. A review of large language models in medical education, clinical decision support, and healthcare administration. *Healthcare (Basel).* (2025) 13(6):603. doi: 10.3390/healthcare13060603
21. Kisvarday S, Yan A, Yarahuan J, Kats DJ, Ray M, Kim E, et al. ChatGPT use among pediatric health care providers: cross-sectional survey study. *JMIR Form Res.* (2024) 8:e56797. doi: 10.2196/56797
22. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J Biomed Inform.* (2024) 151:104620. doi: 10.1016/j.jbi.2024.104620
23. Qarajeh A, Tangpanithandee S, Thongprayoon C, Suppadungsk S, Krisanapan P, Aiumtrakul N, et al. AI-Powered Renal diet support: performance of ChatGPT, Bard AI, and Bing Chat. *Clin Pract.* (2023) 13(5):1160–72. doi: 10.3390/clinpract13050104
24. Aiumtrakul N, Thongprayoon C, Suppadungsk S, Krisanapan P, Miao J, Qureshi F, et al. Navigating the landscape of personalized medicine: the relevance of ChatGPT, BingChat, and Bard AI in nephrology literature searches. *J Pers Med.* (2023) 13(10):1457. doi: 10.3390/jpm13101457
25. OpenAI. Introducing ChatGPT: OpenAI (2022). Available online at: <https://openai.com/blog/chatgpt> (Accessed December 7, 2025).
26. Teperikidis L, Boulmpou A, Papadopoulos C, Biondi-Zoccai G. Using ChatGPT to perform a systematic review: a tutorial. *Minerva Cardiol Angiol.* (2024) 72(6):547–67. doi: 10.23736/S2724-5683.24.06568-2
27. Cheng A, Calhoun A, Reedy G. Artificial intelligence-assisted academic writing: recommendations for ethical use. *Adv Simul (Lond).* (2025) 10(1):22. doi: 10.1186/s41077-025-00350-6
28. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell.* (2023) 6:1169595. doi: 10.3389/frai.2023.1169595
29. Miao J, Thongprayoon C, Cheungpasitporn W. Assessing the accuracy of ChatGPT on core questions in glomerular disease. *Kidney Int Rep.* (2023) 8(8):1657–9. doi: 10.1016/j.ekir.2023.05.014
30. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. *J Med Internet Res.* (2024) 26:e22769. doi: 10.2196/22769
31. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med.* (2023) 388(13):1233–9. doi: 10.1056/NEJMsr2214184
32. Fatima A, Shafique MA, Alam K, Fadlalla Ahmed TK, Mustafa MS. ChatGPT in medicine: a cross-disciplinary systematic review of ChatGPT's (artificial intelligence) role in research, clinical practice, education, and patient interaction. *Medicine (Baltimore).* (2024) 103(32):e39250. doi: 10.1097/MD.00000000000039250
33. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief Bioinform.* (2023) 25(1). doi: 10.1093/bib/bbad493
34. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care.* (2023) 27(1):75. doi: 10.1186/s13054-023-04380-2
35. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA.* (2025) 333(4):319–28. doi: 10.1001/jama.2024.21700
36. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res.* (2023) 25:e48568. doi: 10.2196/48568
37. Park YJ, Pillai A, Deng J, Guo E, Gupta M, Paget M, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Med Inform Decis Mak.* (2024) 24(1):72. doi: 10.1186/s12911-024-02459-6
38. U.S. Food and Drug Administration. U.S. Food and Drug Administration website: U.S. Food and Drug Administration. Available online at: <https://www.fda.gov> (Accessed December 7, 2025).
39. Kidney Disease: Improving Global Outcomes (KDIGO). KDIGO Guidelines: KDIGO (Kidney Disease: Improving Global Outcomes). Available online at: <https://kdigo.org/guidelines/> (Accessed December 7, 2025).
40. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med.* (2025) 31(1):60–9. doi: 10.1038/s41591-024-03425-5
41. Eckardt KU, Agarwal R, Aswad A, Awad A, Block GA, Bacci MR, et al. Safety and efficacy of vadadustat for Anemia in patients undergoing dialysis. *N Engl J Med.* (2021) 384(17):1601–12. doi: 10.1056/NEJMoa2025956
42. Kidney Disease: Improving Global Outcomes (KDIGO). Chapter 1: definition and classification of CKD. *Kidney Int Suppl (2011).* (2013) 3(1):19–62. doi: 10.1038/kisup.2012.64
43. Hellmich B, Sanchez-Alamo B, Schirmer JH, Berti A, Blockmans D, Cid MC, et al. EULAR Recommendations for the management of ANCA-associated vasculitis: 2022 update. *Ann Rheum Dis.* (2024) 83(1):30–47. doi: 10.1136/ard-2022-223764
44. Lieske JC, Ariceta G, Groothoff JW, Lipkin G, Mochhala SH, Schalk G, et al. PHYOX3: nedosiran long-term safety and efficacy in patients with primary hyperoxaluria type 1. *Kidney Int Rep.* (2025) 10(6):1993–2002. doi: 10.1016/j.ekir.2025.03.031
45. Baum MA, Langman C, Cochat P, Lieske JC, Mochhala SH, Hamamoto S, et al. PHYOX2: a pivotal randomized study of nedosiran in primary hyperoxaluria type 1 or 2. *Kidney Int.* (2023) 103(1):207–17. doi: 10.1016/j.kint.2022.07.025
46. Pergola PE, Rosenbaum DP, Yang Y, Chertow GM. A randomized trial of tenapanor and phosphate binders as a dual-mechanism treatment for hyperphosphatemia in patients on maintenance dialysis (AMPLIFY). *J Am Soc Nephrol.* (2021) 32(6):1465–73. doi: 10.1681/ASN.2020101398
47. Matsumura Y. Risk analysis of ecuzimab-related meningococcal disease in Japan using the Japanese adverse drug event report database. *Drug Healthc Patient Saf.* (2020) 12:207–15. doi: 10.2147/DHPS.S257009
48. Abdelgadir Y, Thongprayoon C, Miao J, Suppadungsk S, Pham JH, Mao MA, et al. AI integration in nephrology: evaluating ChatGPT for accurate ICD-10 documentation and coding. *Front Artif Intell.* (2024) 7:1457586. doi: 10.3389/frai.2024.1457586
49. Ho CN, Tian T, Ayers AT, Aaron RE, Phillips V, Wolf RM, et al. Qualitative metrics from the biomedical literature for evaluating large language models in clinical decision-making: a narrative review. *BMC Med Inform Decis Mak.* (2024) 24(1):357. doi: 10.1186/s12911-024-02757-z
50. Naliyathaliyazhayil P, Muthyala R, Gichoya JW, Purkayastha S. Evaluating the reasoning capabilities of large language models for medical coding and hospital readmission risk stratification: zero-shot prompting approach. *J Med Internet Res.* (2025) 27:e74142. doi: 10.2196/74142
51. Tripathi S, Sukumaran R, Cook TS. Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. *J Am Med Inform Assoc.* (2024) 31(6):1436–40. doi: 10.1093/jamia/ocad258
52. Sun QW, Miller J, Hull SC. Charting the ethical landscape of generative AI-augmented clinical documentation. *J Med Ethics.* (2025). doi: 10.1136/jme-2024-110656
53. Tangri N, Cheungpasitporn W, Crittenden SD, Fornoni A, Peralta CA, Singh K, et al. Responsible use of artificial intelligence to improve kidney care: a statement from the American society of nephrology. *J Am Soc Nephrol.* (2025). doi: 10.1681/ASN.00000000929
54. Cheungpasitporn W, Athavale A, Ghazi L, Kashani KB, Colicchio T, Koynier JL, et al. Transforming nephrology through artificial intelligence: a state-of-the-art roadmap for clinical integration. *Clin Kidney J.* (2026) 19(2):sfag004. doi: 10.1093/cjk/sfag004