Original Paper

Exploring the Potential of Claude 3 Opus in Renal Pathological Diagnosis: Performance Evaluation

Xingyuan Li[1*], MS; Ke Liu[1*], MS; Yanlin Lang[1*], MS; Zhonglin Chai[2], Prof Dr; Fang Liu[1], Prof Dr Med

[1]Department of Nephrology, West China Hospital, Sichuan University, Chengdu, China

[2]Department of Diabetes, Central Clinical School, Monash University, Melbourne, Australia

[*]these authors contributed equally

Corresponding Author:
Fang Liu, Prof Dr Med
Department of Nephrology
West China Hospital
Sichuan University
37 Guoxue Alley
Wuhou District
Chengdu, 610041
China
Phone: 86 13890790651
Email: liufangfh@163.com

Abstract

Background: Artificial intelligence (AI) has shown great promise in assisting medical diagnosis, but its application in renal pathology remains limited.

Objective: We evaluated the performance of an advanced AI language model, Claude 3 Opus (Anthropic), in generating diagnostic descriptions for renal pathological images.

Methods: We carefully curated a dataset of 100 representative renal pathological images from the *Diagnostic Atlas of Renal Pathology* (3rd edition). The image selection aimed to cover a wide spectrum of common renal diseases, ensuring a balanced and comprehensive dataset. Claude 3 Opus generated diagnostic descriptions for each image, which were scored by 2 pathologists on clinical relevance, accuracy, fluency, completeness, and overall value.

Results: Claude 3 Opus achieved a high mean score in language fluency (3.86) but lower scores in clinical relevance (1.75), accuracy (1.55), completeness (2.01), and overall value (1.75). Performance varied across disease types. Interrater agreement was substantial for relevance (κ=0.627) and overall value (κ=0.589) and moderate for accuracy (κ=0.485) and completeness (κ=0.458).

Conclusions: Claude 3 Opus shows potential in generating fluent renal pathology descriptions but needs improvement in accuracy and clinical value. The AI's performance varied across disease types. Addressing the limitations of single-source data and incorporating comparative analyses with other AI approaches are essential steps for future research. Further optimization and validation are needed for clinical applications.

(JMIR Med Inform 2024;12:e65033) doi: 10.2196/65033

KEYWORDS

artificial intelligence; Claude 3 Opus; renal pathology; diagnostic performance; large language model; LLM; performance evaluation; medical diagnosis; AI language model; diagnosis; pathology images; pathologist; clinical relevance; accuracy; language fluency; pathological diagnosis

Introduction

Artificial intelligence (AI) has demonstrated remarkable capabilities in analyzing complex medical data and assisting clinical decision-making across various fields [1]. In particular, AI's potential for interpreting histopathological images has been increasingly recognized, offering novel insights into disease pathogenesis and diagnosis [2]. However, the application of AI in renal pathology, a field characterized by high complexity and variability, remains relatively unexplored. Recent advancements

XSL•FO
RenderX

in natural language processing, such as the development of large language models (LLMs) like GPT-3 (OpenAI) and Claude 3 Opus (Anthropic), have opened up new possibilities for AI-assisted pathological diagnosis [3,4]. These models can capture semantic and contextual information from textual data and generate coherent, human-like responses. Despite the proven utility of AI in other pathology domains, such as oncology and dermatology, its performance and feasibility in renal pathology have not been systematically evaluated. To address this gap, we conducted a pioneering study to investigate the potential of using Claude 3 Opus, a state-of-the-art AI language model, for renal pathological diagnosis. By assessing the model's ability to generate accurate and clinically relevant diagnostic descriptions for a wide range of renal pathologies, we aimed to provide initial evidence and insights into the strengths and limitations of AI in this challenging field.

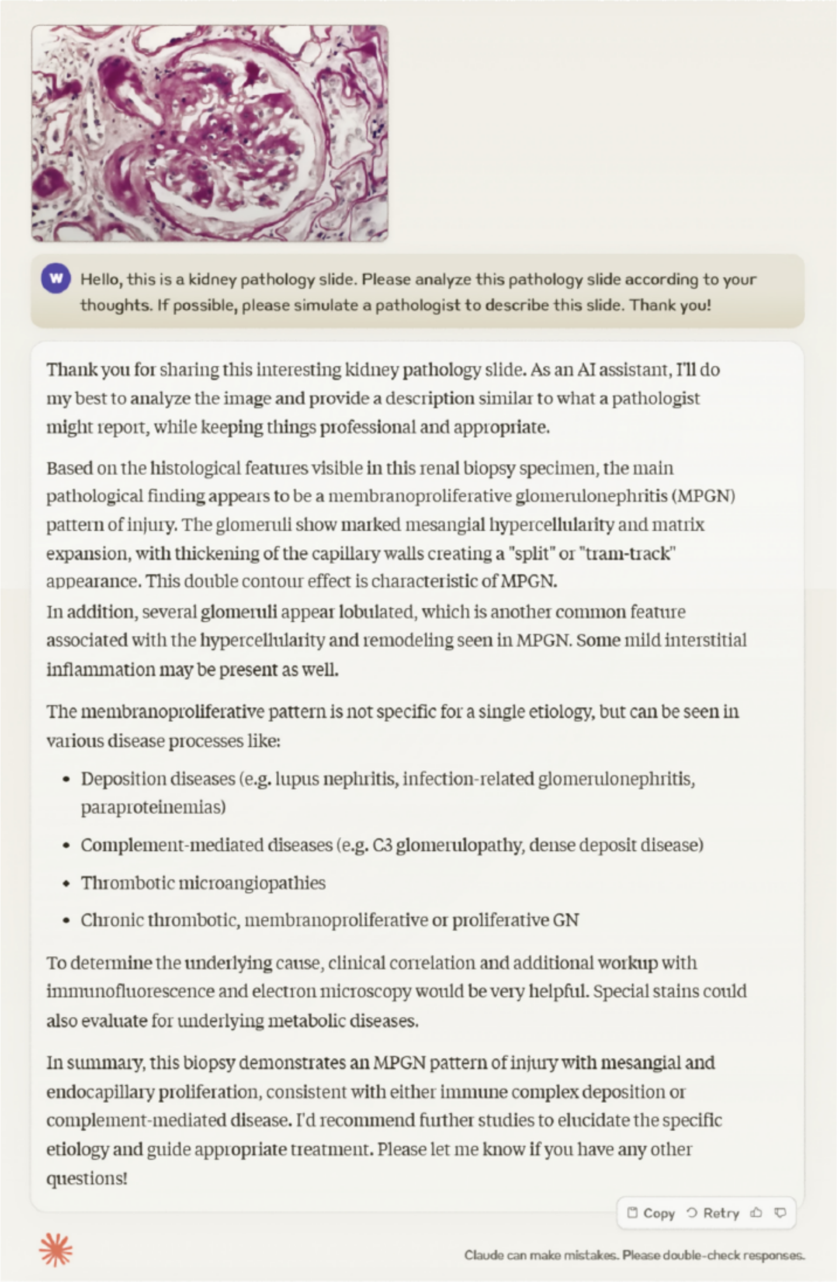
Methods

We carefully curated a dataset of 100 representative renal pathological images from the *Diagnostic Atlas of Renal Pathology* (3rd edition) [5]. The image selection process aimed to cover a wide spectrum of common renal diseases, ensuring a balanced and comprehensive dataset. The number of images per disease type ranged from 2 to 9 (eg, immunoglobulin A nephropathy: n=5; acute tubular injury: n=4; diabetic nephropathy: n=9), proportional to the relative prevalence and morphological diversity of each condition. Multimedia Appendix 1 provides the complete list of images and their corresponding disease labels.

Claude 3 Opus, an advanced AI language model, was used to generate diagnostic descriptions for each image. For each image, the model was given the following prompt: "Describe the key morphological features and provide a diagnostic impression for this renal biopsy image." No additional disease-specific information or background knowledge was provided, allowing us to assess Claude 3 Opus's standalone performance in renal pathology interpretation. The generated descriptions were evaluated by 2 experienced renal pathologists using a comprehensive 5-point scale across 5 key dimensions: clinical relevance, accuracy, language fluency, detail completeness, and overall clinical value. These evaluation dimensions were selected based on established frameworks for assessing the quality and utility of pathology reports [6,7]. They collectively cover the essential aspects of an effective pathology description, from clinical applicability and correctness to clarity and comprehensiveness. Descriptive statistics (means and SDs) of the evaluation scores were calculated for each pathological type. The interrater agreement between the 2 pathologists' scores was assessed using the Cohen κ statistic. Figure 1 shows an example of a prompt to Claude 3 Opus and its response.

XSL•FO
RenderX

Figure 1. Example of a prompt to Claude 3 Opus and its response.



Hello, this is a kidney pathology slide. Please analyze this pathology slide according to your thoughts. If possible, please simulate a pathologist to describe this slide. Thank you!

Thank you for sharing this interesting kidney pathology slide. As an AI assistant, I'll do my best to analyze the image and provide a description similar to what a pathologist might report, while keeping things professional and appropriate.

Based on the histological features visible in this renal biopsy specimen, the main pathological finding appears to be a membranoproliferative glomerulonephritis (MPGN) pattern of injury. The glomeruli show marked mesangial hypercellularity and matrix expansion, with thickening of the capillary walls creating a "split" or "tram-track" appearance. This double contour effect is characteristic of MPGN.

In addition, several glomeruli appear lobulated, which is another common feature associated with the hypercellularity and remodeling seen in MPGN. Some mild interstitial inflammation may be present as well.

The membranoproliferative pattern is not specific for a single etiology, but can be seen in various disease processes like:

- Deposition diseases (e.g. lupus nephritis, infection-related glomerulonephritis, paraproteinemias)
- Complement-mediated diseases (e.g. C3 glomerulopathy, dense deposit disease)
- Thrombotic microangiopathies
- Chronic thrombotic, membranoproliferative or proliferative GN

To determine the underlying cause, clinical correlation and additional workup with immunofluorescence and electron microscopy would be very helpful. Special stains could also evaluate for underlying metabolic diseases.

In summary, this biopsy demonstrates an MPGN pattern of injury with mesangial and endocapillary proliferation, consistent with either immune complex deposition or complement-mediated disease. I'd recommend further studies to elucidate the specific etiology and guide appropriate treatment. Please let me know if you have any other questions!

Claude can make mistakes. Please double-check responses.

Results

The performance evaluation results for Claude 3 Opus in generating renal pathological descriptions are presented in Table 1. The AI model achieved a high overall score in language fluency (mean score 3.86, SD 0.68), indicating its ability to produce grammatically correct and easily readable reports. However, the model's performance in other key aspects was suboptimal, with lower scores for clinical relevance (mean score 1.75, SD 0.77), accuracy (mean score 1.55, SD 0.66), detail completeness (mean score 2.01, SD 0.84), and overall clinical value (mean score 1.75, SD 0.74). Notably, the AI model's performance varied across different renal pathological types. Higher scores (>4) were observed for certain diseases, such as membranoproliferative glomerulonephritis and subacute bacterial endocarditis–associated glomerulonephritis, suggesting the model's potential in assisting the diagnosis of these specific conditions. Conversely, the model's performance was subpar for several other types, including acute interstitial nephritis (mean score 1.00, SD 0.00) and collapsing glomerulopathy (mean score 1.17, SD 0.24), indicating limitations in capturing their key diagnostic features. The inter-rater agreement analysis revealed substantial agreement between the 2 pathologists on clinical relevance (κ=0.627) and overall clinical value (κ=0.589), as well as moderate agreement on accuracy (κ=0.485) and detail completeness (κ=0.458).

XSL•FO
RenderX

Table 1. Performance of Claude 3 Opus in generating descriptions for 100 renal pathological images across 27 disease types.

Standard pathological diagnosis	Clinical relevance, mean score	Accuracy of description, mean score	Language fluency, mean score	Detail completeness, mean score	Overall clinical applicability, mean score
Immunoglobin A nephropathy (n=5)	2.10	1.10	4.30	2.90	1.90
Immunoglobin G4–related tubulointerstitial nephritis (n=3)	1.67	1.17	4.33	2.50	1.67
Proliferative glomerulonephritis with monoclonal deposits (n=3)	2.33	1.83	4.17	2.83	2.50
Autosomal dominant polycystic kidney disease (n=2)	1.25	1.50	3.75	2.50	1.75
Henoch-Schönlein purpura nephritis (n=5)	1.10	1.10	3.90	2.00	1.40
Acute postinfectious glomerulonephritis (n=4)	1.63	1.38	3.88	2.38	1.63
Acute interstitial nephritis (n=3)	1.17	1.00	3.67	1.50	1.00
Acute tubular injury (n=4)	1.25	1.38	3.50	1.88	1.38
Acute pyelonephritis (n=4)	1.38	1.38	3.88	2.38	1.63
Focal segmental glomerulosclerosis (n=4)	1.63	1.75	3.75	2.25	1.88
Anti–glomerular basement membrane antibody–mediated glomerulonephritis (n=4)	2.38	2.38	4.00	2.75	2.63
Chronic interstitial fibrosis and tubular atrophy (n=2)	1.50	1.00	4.00	2.00	2.00
Chronic pyelonephritis (n=2)	2.00	1.50	4.00	2.00	2.00
Diffuse mesangial sclerosis (n=3)	1.83	1.50	3.50	1.50	1.83
Membranoproliferative glomerulonephritis (n=3)	4.17	4.00	3.83	3.67	4.00
Arterionephrosclerosis (n=4)	1.13	1.00	4.00	1.00	1.13
Medullary cystic disease (n=3)	1.33	1.17	3.83	1.67	1.33
Collapsing glomerulopathy (n=3)	1.17	1.17	3.50	1.17	1.17
Diabetic nephropathy (n=9)	1.50	1.44	3.61	1.50	1.44
Preeclampsia (n=8)	1.44	1.56	3.75	1.94	1.69
Fibrillary glomerulonephritis (n=2)	1.50	1.25	3.75	1.50	1.25
Microscopic polyangiitis (n=3)	1.83	1.67	3.83	1.67	1.83
Hemoglobinuric acute renal failure (n=2)	2.25	1.75	4.00	2.75	2.25
Thrombotic microangiopathy (n=3)	1.50	1.50	4.00	1.67	1.67
Subacute bacterial endocarditis–associated glomerulonephritis (n=3)	3.17	3.00	4.00	2.67	2.83
Scleroderma (n=4)	2.00	1.88	4.00	1.75	1.88
Kidney biopsies from healthy people (n=5)	1.50	1.10	3.50	1.40	1.30
Overall	1.75	1.55	3.86	2.01	1.75

Discussion

Principal Findings

This study provides initial evidence for the potential of advanced AI language models, such as Claude 3 Opus, in assisting renal pathological diagnosis. The model demonstrated promise in generating fluent and readable pathological descriptions, which could streamline the reporting process and alleviate pathologists' workloads. However, the suboptimal performance in accuracy, clinical relevance, and overall value highlights the need for further improvement before clinical implementation. The interrater agreement analysis revealed substantial agreement for clinical relevance and overall value, but only moderate agreement for accuracy and completeness. This discrepancy might stem from the inherent subjectivity in evaluating granular aspects of pathology descriptions. Pathologists' individual expertise, expectations, and interpretive styles could influence their assessments of accuracy and completeness. Developing standardized scoring rubrics and involving larger, multicenter expert panels in future studies could help mitigate this variability and improve evaluation reliability [8,9].

XSL•FO
RenderX

The AI model's performance varied notably across different renal pathological types. The higher scores for conditions like membranoproliferative glomerulonephritis and infection-related glomerulonephritis could be attributed to their distinct morphological features, such as characteristic immune-complex deposits or structural alterations [10]. These overt patterns might be more readily discernible by the AI algorithms. Conversely, the lower performance for diseases like acute interstitial nephritis and collapsing glomerulopathy might reflect their subtler or more heterogeneous histological manifestations [11], posing challenges for automated interpretation.

While Claude 3 Opus exhibited potential in generating fluent descriptions, the limited accuracy and clinical relevance underscore the challenges of applying LLMs to complex medical image data. As highlighted by recent studies [12,13], LLMs excel at processing textual information but may struggle with the intricacies of specialized visual tasks like histopathology interpretation. Continued research on architectures and training strategies tailored for medical vision applications is crucial for realizing the full potential of AI in this domain.

Our study's reliance on images from a single atlas dataset may have introduced some biases and limits the generalizability to real-world clinical scenarios. Although the *Diagnostic Atlas of Renal Pathology* is widely recognized as a high-quality reference, external validation using diverse, multicenter biopsy datasets is essential to assess the robustness and transferability of our findings [14]. Future studies should prioritize prospective validation on independent clinical cohorts to establish the real-world performance of AI models like Claude 3 Opus.

Comparing the performance of Claude 3 Opus with other AI models or traditional diagnostic methods could offer valuable insights into its relative strengths and areas for improvement. While direct comparisons were beyond the scope of this initial study, recent work has reported promising results with deep learning–based approaches for renal pathology classification. Shen et al [15] developed a convolutional neural network (CNN) model that achieved an accuracy of 87.5% in classifying 6 common glomerular diseases. Similarly, Hermsen et al [16] demonstrated the effectiveness of a multiclass CNN for diagnosing various renal pathologies, with an area under the receiver operating characteristic curve ranging from 0.88 to 0.99. These studies highlight the potential of specialized AI architectures for renal pathology diagnosis, serving as benchmarks for evaluating the performance of LLMs like Claude 3 Opus. Future research should aim to conduct head-to-head comparisons and explore synergistic integrations of LLMs with image-based AI models to leverage their complementary strengths.

To fully realize AI's potential in renal pathology, future research should focus on optimizing the model's training process with comprehensive and balanced datasets, incorporating expert feedback into the learning process, and integrating AI with advanced digital pathology tools for more accurate and objective diagnoses. The scarcity of large-scale annotated datasets and concerns about AI interpretability remain major challenges to be addressed. Amidst these challenges, the collaboration between AI researchers, pathologists, and clinicians is crucial for developing reliable and clinically applicable AI models. Standardized evaluation frameworks and best practices for the responsible use of AI in pathology are also needed. As AI continues to advance, we anticipate its increasing role in enhancing diagnostic accuracy and efficiency, ultimately benefiting patient care.

Conclusion

In conclusion, our study provides an initial assessment of Claude 3 Opus's potential for AI-assisted renal pathology diagnosis. While the model showed promise in generating fluent descriptions, improvements in accuracy and clinical relevance are necessary for practical implementation. The observed performance variations across disease types highlight the need for targeted model optimizations. Addressing the limitations of single-source data and incorporating comparative analyses with other AI approaches are essential steps for future research. As AI continues to advance, close collaboration between pathologists, AI researchers, and clinicians will be instrumental in developing reliable, integrated diagnostic solutions that enhance patient care in renal pathology.

Acknowledgments

This study was supported by Health Commission of Sichuan Province Program (21ZD001)

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Complete list of images and their corresponding disease labels.
[PDF File (Adobe PDF File), 213 KB-Multimedia Appendix 1]

References

XSL•FO
RenderX

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. Jan 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
2. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol. May 2019;20(5):e253-e261. [FREE Full text] [doi: 10.1016/S1470-2045(19)30154-8] [Medline: 31044723]
3. Claude: A breakthrough AI assistant based on constitutional AI. Anthropic. URL: https://www.anthropic.com/claude.html [accessed 2023-05-15]
4. Brown T, Mann B, Ryder N. Language models are few-shot learners. arXiv. Preprint posted online May 28, 2020. [doi: 10.48550/arXiv.2005.14165]
5. Fogo A, Kashgarian M. Diagnostic Atlas of Renal Pathology, 3rd edition. Amsterdam, Netherlands. Elsevier; 2016.
6. Hewitt KJ, Wiest IC, Carrero ZI, Bejan L, Millner TO, Brandner S, et al. Large language models as a diagnostic support tool in neuropathology. J Pathol Clin Res. Nov 2024;10(6):e70009. [FREE Full text] [doi: 10.1002/2056-4538.70009] [Medline: 39505569]
7. Loor-Torres R, Wu Y, Esteban Cabezas, Borras-Osorio M, Toro-Tobon D, Duran M, et al. Use of natural language processing to extract and classify papillary thyroid cancer features from surgical pathology reports. Endocr Pract. Nov 2024;30(11):1051-1058. [doi: 10.1016/j.eprac.2024.08.008] [Medline: 39197747]
8. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. Jan 2016;69:245-247. [doi: 10.1016/j.jclinepi.2015.04.005] [Medline: 25981519]
9. Liu X, Duan C, Kim M, Zhang L, Jee E, Maharjan B, et al. Claude 3 Opus and ChatGPT with GPT-4 in dermoscopic image analysis for melanoma diagnosis: comparative performance analysis. JMIR Med Inform. Aug 06, 2024;12:e59273. [FREE Full text] [doi: 10.2196/59273] [Medline: 39106482]
10. Sethi S, Haas M, Markowitz G, D'Agati VD, Rennke HG, Jennette JC, et al. Mayo Clinic/Renal Pathology Society consensus report on pathologic classification, diagnosis, and reporting of GN. J Am Soc Nephrol. May 2016;27(5):1278-1287. [FREE Full text] [doi: 10.1681/ASN.2015060612] [Medline: 26567243]
11. Mubarak M. Collapsing focal segmental glomerulosclerosis: increasing the awareness. J Nephropathol. Jul 2012;1(2):77-80. [FREE Full text] [doi: 10.5812/nephropathol.7474] [Medline: 24475392]
12. Sarangi PK, Irodi A, Panda S, Nayak DSK, Mondal H. Radiological differential diagnoses based on cardiovascular and thoracic imaging patterns: perspectives of four large language models. Indian J Radiol Imaging. Apr 2024;34(2):269-275. [FREE Full text] [doi: 10.1055/s-0043-1777289] [Medline: 38549881]
13. Nowak S, Schneider H, Layer YC, Theis M, Biesner D, Block W, et al. Development of image-based decision support systems utilizing information extracted from radiological free-text report databases with text-based transformers. Eur Radiol. May 2024;34(5):2895-2904. [FREE Full text] [doi: 10.1007/s00330-023-10373-0] [Medline: 37934243]
14. Subramanian S, Viswanathan VK, Ramani S, Rajendiran P. Generalizability in the age of deep learning: The case of renal histology. Kidney Int Rep. 2022;7(8):1623-1625. [FREE Full text] [doi: 10.1016/j.ekir.2022.05.011]
15. Shen L, Sun W, Zhang Q, Wei M, Xu H, Luo X, et al. Deep learning-based model significantly improves diagnostic performance for assessing renal histopathology in lupus glomerulonephritis. Kidney Dis (Basel). Jul 2022;8(4):347-356. [FREE Full text] [doi: 10.1159/000524880] [Medline: 36157261]
16. Hermsen M, de Bel T, den Boer M, Steenbergen EJ, Kers J, Florquin S, et al. Deep learning-based histopathologic assessment of kidney tissue. J Am Soc Nephrol. Oct 2019;30(10):1968-1979. [FREE Full text] [doi: 10.1681/ASN.2019020144] [Medline: 31488607]

Abbreviations

AI: artificial intelligence
CNN: convolutional neural network
LLM: large language model

Edited by C Lovis; submitted 02.08.24; peer-reviewed by X Liu, PK Sarangi, GK Gupta; comments to author 22.09.24; revised version received 28.09.24; accepted 19.10.24; published 15.11.24

Please cite as:
Li X, Liu K, Lang Y, Chai Z, Liu F
Exploring the Potential of Claude 3 Opus in Renal Pathological Diagnosis: Performance Evaluation
JMIR Med Inform 2024;12:e65033
URL: https://medinform.jmir.org/2024/1/e65033
doi: 10.2196/65033
PMID:

XSL•FO
RenderX

©Xingyuan Li, Ke Liu, Yanlin Lang, Zhonglin Chai, Fang Liu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 15.11.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX