Open Access
Poster

Cureus

Transforming Clinical Practice with AI: Evaluating Large Language Models in Complex Neurological Care

Open Access
Published 04/04/2025

Skylar Bentley [1]

1. Research, Orlando College of Osteopathic Medicine, Winter Garden, USA

Corresponding author: Skylar Bentley, sbentley@students.ocom.org

Categories: Other
Keywords: ai in healthcare, artificial intelligence (ai), clinical decision support, diagnosis and treatment planning, guillain-barré syndrome (gbs), large language models (llms), machine learning in medicine, neurology

How to cite this poster
Bentley S (2025) Transforming Clinical Practice with AI: Evaluating Large Language Models in Complex Neurological Care. Cureus 17(4): e.

Abstract

Background and Introduction:
AI is transforming clinical workflows, providing decision support for complex medical cases. Large Language Models (LLMs) like ChatGPT, Google Gemini, and Claude 3.5 Sonnet have demonstrated diagnostic potential but require further evaluation in multifaceted clinical presentations. GBS with SEL presents a unique diagnostic and therapeutic challenge, demanding precise evaluation, timely intervention, and comprehensive follow-up. This study examines whether AI-generated treatment plans align with expert physician recommendations in managing this.

Objective:

1. Evaluate the accuracy of AI-generated treatment plans for Guillain-Barré Syndrome (GBS) with Spinal Epidural Lipomatosis (SEL).
2. Compare AI-generated management plans (ChatGPT, Google Gemini, Claude 3.5 Sonnet) to expert physician recommendations.
3. Assess AI performance in diagnostic reasoning, treatment selection, and follow-up planning.
4. Determine the clinical reliability of LLMs as decision-support tools in neurology.
5. Identify limitations in AI-generated plans and areas for improvement in long-term patient management.
6. Explore the potential for AI integration into clinical workflows while ensuring physician oversight.

Methods:

1. Case Input: A structured GBS with SEL case was entered into ChatGPT, Google Gemini, and Claude 3.5 Sonnet to generate treatment plans.
2. AI Outputs: Models provided diagnoses, treatment plans (IVIG, plasmapheresis), and follow-up strategies.
3. Evaluation: Four board-certified physicians assessed outputs using a 20-point rubric. Blinded review ensured unbiased scoring across diagnosis, treatment, and follow-up.
4. Statistical Analysis: ANOVA ($p < 0.05$) tested significance between models. This study provides a rigorous, unbiased comparison of AI in complex neurological care.

Results

1. Overall Performance: LLMs generated clinically sound GBS treatment plans. Claude 3.5 Sonnet (18.5/20) outperformed ChatGPT (17.5/20) and Google Gemini (17.25/20). ANOVA ($F = 1.36$, $p = 0.29$) showed no significant differences in overall accuracy.
2. Diagnosis: All models correctly identified GBS, with Claude 3.5 Sonnet providing the most detailed reasoning.
3. Treatment: AI recommended IVIG, plasmapheresis, and monitoring effectively, closely aligning with expert care.

4. Follow-Up: Physician: 5/5 (comprehensive rehab & follow-up); Claude 3.5 Sonnet: 4.5/5 (minor rehab gaps).ChatGPT & Gemini: 4/5 (adequate but lacked detail).
5. Key Trend: AI models were strongest in diagnosis and treatment selection but less detailed in follow-up planning.

Conclusion:
LMs show promise as decision-support tools but are not substitutes for clinical expertise. Claude 3.5 Sonnet performed best, yet all models required physician validation for long-term care planning. Future Directions: Refining AI-generated follow-up strategies, integrating real-world clinical data, and optimizing AI-assisted workflows to complement expert decision-making. With further refinement, AI could become a valuable asset in neurological care, improving efficiency, accuracy, and patient outcomes.