Journal of Pain Research

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Smart Processing and Intelligent Navigation for Evaluation (SPINE): Comparing Clinicians and AI Language Model (GPT-4) in Spinal Cord Stimulation Candidate Selection

Giuliano Lo Bianco[1,*], Alexandra Therond[2,*], Francesco Paolo D'angelo[3], Leonardo Kapural[4], Sudhir Diwan[5], Peter Staats[6,7], Sean Li[8], Paul J Christo[9], Timothy R Deer[10], Christopher L Robinson[9]

[1]Anesthesiology and Pain Department, Fondazione Istituto G. Giglio Cefalù, Palermo, Italy; [2]Department of Psychology, Université du Québec à Montréal, Montréal, QC, Canada; [3]Department of Anaesthesia, Intensive Care and Emergency, University Hospital Policlinico Paolo Giaccone, Palermo, Italy; [4]Center for Clinical Research, Carolinas Pain Institute, Winston-Salem, NC, USA; [5]Albert Einstein College of Medicine, Bronx, NY, USA; [6]electroCore, Rockaway, NJ, USA; [7]National Spine and Pain Centers, Rockville, MD, USA; [8]National Spine and Pain Centers, Shrewsbury, NJ, USA; [9]Division of Pain Medicine, Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA; [10]The Spine and Nerve Centers of the Virginias, Charleston, WV, USA

*These authors contributed equally to this work

Correspondence: Giuliano Lo Bianco, Anesthesiology and Pain Department, Fondazione Istituto G. Giglio Cefalù, Palermo, Italy, Email giulianolobianco@gmail.com; Christopher L Robinson, Johns Hopkins University School of Medicine, Department of Anesthesiology and Critical Care Medicine, 1800 Orleans Street, Baltimore, MD, 21287, USA, Email ChristopherRobinsonMDPhD@outlook.com

Background: With the continued advancement of artificial intelligence (AI), large language models (LLMs) such as GPT-4 may assist clinicians in evaluating patient candidacy for spinal cord stimulation (SCS). We compared a general-purpose, non–fine-tuned LLM (GPT-4), an expert multidisciplinary team (MDT), and a clinician-input, rule-based e-Health decision-support tool. The study focused exclusively on decision agreement and did not assess clinical outcomes (eg, pain relief or device retention).

Methods: This single-center, retrospective cohort was conducted at Fondazione Istituto G. Giglio (Cefalù, Italy) and included 93 consecutive adults referred to the MDT for SCS evaluation between January 2022 and March 2024. The MDT issued binary recommendations ("proceed" vs "do not proceed") as the reference standard. The e-Health tool generated "yes", "maybe", or "no" outputs from structured clinician-entered data. GPT-4 was applied zero-shot, using a single standardized prompt on anonymized vignettes within an offline environment. The primary endpoint was agreement (weighted κ) among MDT, e-Health, and GPT-4; sensitivity/specificity analyses explored three interpretations of "maybe".

Results: The MDT recommended SCS for 91.4% of patients, compared with 54.8% for the e-Health tool and 46.2% for GPT-4. Agreement was moderate for MDT vs e-Health (κ = 0.51) and e-Health vs GPT-4 (κ = 0.46), and fair for MDT vs GPT-4 (κ = 0.29). GPT-4 demonstrated a more conservative profile, favoring specificity over sensitivity.

Conclusion: A non–fine-tuned GPT-4 approximated but did not replicate MDT decision-making, functioning as a high-specificity, low-sensitivity filter. A layered workflow combining rule-based tools with expert oversight and targeted LLM adaptation may best optimize SCS candidate selection.

Keywords: artificial intelligence, large-language models, spinal cord stimulation, chronic pain, patient selection, neuromodulation

Background

Artificial intelligence (AI) applications in the clinical setting have made significant progress, from early-stage research prototypes to becoming an everyday clinical asset, accelerating information retrieval, and pattern recognition across almost every medical specialty.[1–5] Large-language models (LLMs) such as GPT-4 can now interpret narrative clinical

https://doi.org/10.2147/JPR.S559514
Received: 8 August 2025
Accepted: 13 January 2026
Published: 17 January 2026

vignettes, imaging reports, and laboratory data in a fraction of the time, producing provisional, evidence-linked suggestions.[4,6,7] Nevertheless, the suggestions may include hallucinations, plausible sounding yet fabricated information, an issue occurring even in curated systems, and must therefore be weighed cautiously by clinicians.[4,6,7] Evidence is growing demonstrating that clinicians with the assistance of LLMs (trained and untrained) can solve complex case-management questions more accurately than alone or accelerating documentation such as in the reconciling of medication lists or tailoring peri-operative antibiotic prophylaxis.[4,8–12] Yet in other settings, the same models add little value or misprioritize differential diagnoses, as has been reported in rare metabolic disorders or pediatric dermatology.[8–12] This variability highlights a central informatics challenge – identifying which tasks can be safely augmented by a general-purpose, untrained models, how much clinician oversight is needed, and which require further specialty specific training of the LLM.

Successive waves of fine-tuning and clinician-in-the-loop reinforcement have started to narrow the gap. Trained LLMs evaluated on structured clinical vignettes and simulated clinical encounters, now meet or even exceed average clinician scores.[9,10,13] Evidence from LLMs answering opioid therapy and neuromodulation queries with strong reliability and comprehensibility further illustrates the potential of even untrained models.[6,12,14] LLMs now stand alongside an ecosystem of digital decision-support tools from dermatology image-analysis software and real-time pain-expression detection systems to algorithmic care pathways, such as for neuromodulation, embedded in electronic health records.[7,11,15,16]

Spinal cord stimulation (SCS) therapy is a clear example when optimal patient selection requires the integration of different biomedical and psychosocial information.[17–20] Contemporary consensus statements and e-Health tools provide useful tools yet still leave room for clinical discretion.[16,20–25] We, therefore, explored how an untrained LLM compares with a multidisciplinary team's (MDT) decisions and with an established e-Health tool in evaluating the appropriateness for when SCS therapy is indicated. By clarifying its strengths and limitations, this study offers a preliminary reference for clinicians exploring the integration of AI tools into pain management workflows.[14]

Importantly, the present analysis focuses exclusively on agreement among decision-making systems and does not evaluate downstream clinical outcomes such as pain relief, device retention, or explantation rates. However, current MDT evaluations remain subject to inter-rater variability, are time- and resource-intensive, and may be influenced by subjective bias. Objective AI-based tools could help improve consistency and efficiency in candidate triage.

Methods

The SPINE (Smart Processing and Intelligent Navigation for Evaluation) study is a single-center, retrospective cohort study conducted at the Fondazione Istituto G. Giglio in Cefalù (Palermo, Italy). The study compared three decision-making systems: the multidisciplinary team (MDT, reference standard), a validated rule-based clinician-input e-Health decision-support tool, and an untrained large language model (LLM; GPT-4). Data were analyzed from a prospectively managed clinical "SCS Pathway" study performed by the MDT, as described below. Individuals included from the SCS Pathway study were ≥18 years of age with a chronic pain diagnosis who were referred to the MDT "SCS Pathway" between January 2022 and March 2024 for consideration of SCS therapy. All consecutive adult patients (≥18 years) referred to the MDT during this period were included. No pre-screening beyond incomplete data or missing follow-up was applied, minimizing selection bias. The MDT consists of structured assessments by two pain physicians, a psychologist, and a specialist nurse, who convene to decide candidacy. The MDT issued a binary recommendation ("proceed with SCS trial" or "do not proceed"), which served as the clinical reference standard for subsequent comparisons against the e-Health tool and GPT-4. The e-Health tool generated three possible outputs ("yes", "maybe", and "no"), synthesizing clinical and psychosocial inputs according to predefined weighting rules derived from expert consensus. The e-Health tool (https://scstool.org) is a clinician-input, rule-based algorithm developed from the European consensus recommendations for appropriate referral and selection of patients with chronic pain for SCS.[16,25] It synthesizes predefined clinical and psychosocial parameters into a weighted decision output ("yes", "maybe", "no") reflecting guideline-based thresholds. The algorithm does not use self-reported or machine-learning data but structured clinician-entered variables.

Electronic health-record screening identified 93 patients. The study focused exclusively on agreement metrics across evaluators; clinical outcomes such as pain reduction, implant success, or explant rates were not analyzed. For each patient, demographic variables (age, sex, and body mass index), pain etiology, prior spine surgeries, pain phenotype (neuropathic vs mixed), psychosocial factors (anxiety, depression, and substance dependence), comorbidities, medication profile, and baseline scores for visual analogue scale (VAS), Douleur Neuropathique 4 (DN4), and EuroQol-5 Dimension (EQ-5D) were extracted. Follow-up scores were collected at 3, 6, 12, 18, and 24 months after permanent implantation when applicable. Patients were excluded if they had incomplete follow up.

For the LLM, anonymized case vignettes without identifying information were entered into separate, offline sessions of ChatGPT-4 (OpenAI, March 2024 release). Each vignette included the same variables provided to the MDT and e-Health tool. GPT-4 was applied in a zero-shot manner using a single standardized prompt without temperature variation, role conditioning, or chain-of-thought prompting, to simulate a general-purpose, untrained clinical reasoning context. The standardized question used for all cases was: "Is spinal cord stimulation appropriately indicated for this patient?"

GPT-4's first explicit statement ("yes", "no", or "maybe") determined the categorical output. Each case was processed in an offline, isolated GPT-4 environment (March 2024 release), ensuring no external data retrieval or model training interaction. Since the vignettes were generated after all clinical decisions had been finalized, the LLM and e-Health tool had no access to MDT decisions and played no role in patient care since it was entered retrospectively. Primary endpoints were the level of agreement between LLM, the e-Health tool, and MDT decisions. Institutional Review Board approval was obtained from Comitato Etico Locale Palermo 1 (protocol #23; 19 September 2024). All patient data were fully anonymized prior to processing. No identifiable information was transmitted to third-party servers, and the GPT-4 environment operated locally to ensure compliance with EU GDPR and institutional data-protection policies. The board waived individual informed consent for the retrospective analysis of the anonymized data collected from the prospective study. All procedures conformed to the Declaration of Helsinki and applicable data-protection regulations.

Statistical Analysis

Statistical analyses were conducted using R (version 4.1.2) and the irr package for conducting inter-rater reliability analyses.[26] The weighted kappa (κ) statistic was used to assess the agreement between the MDT, the e-Health tool, and LLM.[27] Pairwise comparisons of the assessment methods were conducted using the formula:

$$K_w = 1 - \frac{\sum^{w_{ij}} O_{ij}}{\sum^{w_{ij}} E_{ij}}$$

where w are the weighting factors, O are the observed frequencies of agreement or disagreement and E are the expected frequencies under the assumption that the assessment methods are independent. For the weighing factors, a value of 1 was used for complete agreement, where both assessment methods either accepted or rejected the patient and a value of 0 used for complete disagreement, where one assessment method accepted the patient while the other rejected them. For partial agreement, such as when one assessment method accepted or rejected the patient while the other provided a tentative decision, squared weights were used, that is, a value of 0.25. This weighting scheme penalized larger disagreements (eg, accepted vs rejected) more heavily than partial ones (eg, accepted vs tentative). Agreement strength was interpreted according to Altman's criteria: κ <0.20 poor, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 good, >0.80 very good.[27]

Results

The cohort included 93 adults with a mean age of 66.1 ± 10.5 years (Table 1). Of the cohort, 40.9% (38/93) were women and 59.1% (55/93) were men. The leading diagnoses were chronic back and leg pain 41.9% (39/93) and post-surgical pain syndrome 36.6% (34/93) with less common diagnoses being neuropathic pain syndrome 11.8% (11/93), complex regional pain syndrome 8.6% (8/93), and idiopathic pain syndrome 1.1% (1/93).

SCS implant decision outcomes varied across the three assessment methods. Of the patients referred to the MDT, SCS therapy was indicated in 91.4% (85/93) (Table 2). In comparison, the e-Health tool and LLM only indicated SCS therapy

Table 1 Demographic and Clinical Characteristics of Study Population

Characteristic	Total (N = 96)
Age in years, M (SD)	66.13 (10.54)
Gender, n (%)	
Male	55 (59.1%)
Female	38 (40.9%)
Chronic pain diagnosis, n (%)	
Chronic back and leg pain	39 (41.9%)
Complex regional pain syndrome	8 (8.6%)
Idiopathic pain syndrome	1 (1.1%)
Neuropathic pain syndrome	11 (11.8%)
Post-surgical pain syndrome	34 (36.6%)
Previous spine surgery, n (%)	
Yes	34 (36.6%)
No	59 (63.4%)
Pain type, n (%)	
Mixed	21 (22.6%)
Neuropathic	72 (77.4%)

Table 2 SCS Implant Decisions Across Assessment Methods

Decision	Multidisciplinary Team n (%)	e-Health Tool n (%)	GPT-4 n (%)
Accepted	85 (91.4%)	51 (54.8%)	43 (46.2%)
Rejected	8 (8.6%)	9 (9.7%)	25 (26.9%)
Tentative	N/A	33 (35.5%)	25 (26.9%)

in 54.8% (51/93) and 46.2% (43/93), respectively. SCS therapy was not indicated in 8.6% (8/93) for the MDT, 9.7% (9/93) for the e-Health tool, and 26.9% (25/93) for LLM. Additionally, both the e-Health decision-support platform and LLM included tentative decisions, accounting for 35.5% (33/93) and 26.9% (25/93), respectively, a category not applicable to the MDT (Table 2).

The MDT and the e-Health tool had the highest rate of complete agreement 58/93 (62.4%) with minimal disagreement 2/93 (2.2%) and moderate partial agreement 35.5% (33/93) (Table 3) with a weighted κ of 0.51, indicating moderate agreement. Comparisons between the MDT and LLM demonstrated a lower rate of complete agreement 51/93 (54.8%) and a higher rate of complete disagreement 17/93 (18.3%), with partial agreement at 25/93 (26.9%) with a weighted κ of 0.29, reflecting fair agreement. Finally, the e-Health tool and LLM achieved complete agreement in 61.3% (57/93), complete disagreement in 8.6% (8/93), and partial agreement in 30.1% (28/93) with a weighted kappa of κ = 0.46, indicating moderate agreement.

A sensitivity and specificity analysis was conducted to evaluate assessments made by the e-Health tool and LLM, using the decision made by the MDT as the reference standard (Table S1). Since specificity and sensitivity rely on binary

Table 3 Pairwise Comparisons of Agreement, Disagreement, Partial Agreement, and Weighted Kappa Values

Assessment Methods Compared		Complete Agreement n (%)	Complete Disagreement n (%)	Partial Agreement n (%)	K_w	Interpretation
Multidisciplinary Team	e-Health tool	58 (62.4%)	2 (2.2%)	33 (35.5%)	0.51	Moderate agreement
Multidisciplinary Team	GPT-4	51 (54.8%)	17 (18.3%)	25 (26.9%)	0.29	Fair agreement
e-Health tool	GPT-4	57 (61.3%)	8 (8.6%)	28 (30.1%)	0.46	Moderate agreement

classification, the analysis treated the tentative responses through three different scenarios. The first scenario excluded tentative decisions, removing uncertainty from the calculation. The second scenario treated tentative decisions as false negatives, if uncertainty leads to rejections. This lowers sensitivity and makes the assessment method appear more conservative. The third scenario treated the tentative decisions as false positives, if uncertainty leads to acceptance. This lowers specificity and makes the assessment method appear more lenient. Overall, the LLM remains more conservative, favoring specificity at the cost of sensitivity, while e-health is more lenient, demonstrating greater sensitivity but slightly lower specificity when tentative decisions are factored in.

Discussion

The present study demonstrates that an untrained LLM can approximate but not yet replicate the nuanced triage decisions of an expert MDT when selecting candidates for SCS. Overall agreement between GPT-4 and MDT was fair ($\kappa = 0.29$) and markedly lower than the moderate concordance observed between MDT and the e-Health tool ($\kappa = 0.51$). The main driver of discordance was GPT-4's conservative stance as SCS therapy was indicated less than half of potential implants 46.2% (43/93) and never indicated SCS therapy when it was not indicated by the MDT, resulting in 100% specificity in each of the three sensitivity-analysis scenarios. The model performed reliably at identifying clear exclusions often patients with substance dependence, mixed pain phenotypes, or scant responses to prior therapy but missed a proportion of individuals who ultimately benefited from implantation. This conservative behavior suggests that GPT-4 may act as a highly specific but less sensitive "filter", an observation consistent with findings from other medical domains where untrained language models tend to avoid over-recommendation when uncertainty is high.

Sensitivity analyses illustrate the trade-off. When "maybe" responses were omitted, GPT-4 had 100% specificity but lost approximately one-quarter of true positives (sensitivity = 0.72). Reclassifying "maybe" as false negatives accentuated this shortfall, whereas regarding them as false positives confirmed GPT4's conservative appearing approach to SCS therapy. This pattern implies that an untrained LLM may serve as a conservative screening adjunct useful for filtering out clear non-candidates but is not yet suitable as a stand-alone gatekeeper where the goal is to maximize access and sensitivity.

The e-Health tool's performance ($\kappa = 0.51$) closely mirrored its original validation study ($\kappa \approx 0.54$) reinforcing the reproducibility of rule-based digital decision aids in SCS candidate selection.[25] Its structured, clinician-input design allowed greater alignment with MDT reasoning than the generative, probabilistic logic of GPT-4. This suggests that model accuracy depends less on computational complexity than on how closely the algorithm's architecture and data mirror the clinical reasoning process.

Our findings echo earlier reports from other fields, where untrained LLMs displayed inconsistent performance on specialized tasks until they underwent targeted, domain-specific training. For instance, trained LLMs have equaled or surpassed clinician accuracy in simulated national licensing examinations, dermatologic image triage, and opioid counselling chatbots, all while improving explanatory transparency and equity across demographic subgroups.[6,9,10,12–14] Machine-learning models specifically developed for SCS outcomes or chronic pain prediction (Karri et al, 2020; Celestin et al, 2009; Russo et al, 2021) have reported accuracies around 70–80%, comparable to the moderate agreement observed here.[28] These findings collectively support the view that algorithmic tools can complement but not yet replace expert multidisciplinary evaluation.

The differential profile of the 25 patients approved by the MDT but declined by GPT-4 further highlights the need for targeted optimization. Compared with cases where SCS was indicated, these individuals more frequently exhibited mixed pain presentations, lower prior treatment gains, and psychosocial comorbidities variables that are semi-structured or narrative-rich and therefore harder for a generic LLM to weigh appropriately. Augmenting the model with structured psychological scales, longitudinal registry data, and examples emphasizing "gray-zone" scenarios could substantially improve discrimination and contextual accuracy.

One methodological limitation concerns the prompting strategy. GPT-4 was evaluated using a single, simple zero-shot prompt without role conditioning or temperature variation. While this design allowed a standardized baseline comparison, it constrained the model's reasoning depth. Future studies should evaluate multi-prompt or chain-of-thought strategies to enhance reproducibility, contextual depth, and transparency. Incorporating rationale-generation (eg, explain-your-answer prompts) would further facilitate interpretability and clinician trust.[14]

Interpretability remains a major concern: although GPT-4's recommendations sometimes aligned with the MDT, no explicit rationale was requested or provided, reinforcing the "black-box" nature of AI systems inputs go in, outputs emerge, yet the decision pathway remains largely opaque.[1,8] Embedding rationale extraction techniques—such as chain-of-thought transparency or rule-based attribution—could bolster clinician trust and facilitate root-cause analysis when clinicians and LLMs.

Finally, it is noteworthy that the e-Health tool, despite its rule-based architecture and absence of generative capabilities, achieved higher agreement with experts than the LLM. These findings suggest that performance depends more on how closely the training data match the clinical task than on the complexity of the model. Future clinical AI systems will likely combine deterministic algorithms that enforce guideline conformity with LLMs capable of contextualizing narrative data creating a hybrid, layered decision-support ecosystem rather than a single autonomous model.

Limitations

Several limitations should be acknowledged. First, this was a single-center study reflecting a specific regional referral pattern, which may not capture variability in SCS candidate assessment across different healthcare systems. Second, the retrospective design depended on the accuracy and completeness of electronic health records, particularly for psychosocial variables, which were not always standardized. Third, the MDT that served as the reference standard is embedded within the same institutional workflow as the e-Health tool, potentially inflating observed agreement due to shared clinical culture.

Fourth, the LLM was evaluated "out-of-the-box", without domain-specific fine-tuning or multi-prompt prompting strategies. Its conservative behavior may therefore reflect prompt simplicity rather than genuine clinical reasoning competence. Fifth, manual post hoc classification of GPT-4 outputs into categorical recommendations ("yes", "no", "maybe") introduced interpreter judgment that may have influenced κ statistics. Sixth, long-term follow-up beyond 24 months was unavailable for a proportion of implanted patients, limiting the ability to correlate model recommendations with sustained clinical benefit or device retention. Finally, the modest sample size precluded subgroup analyses for less frequent pain syndromes and psychological profiles. Future multicenter, prospective studies using standardized psychological and outcome metrics are warranted to validate and refine these findings.

Conclusion

Overall, this study provides an initial benchmark for integrating AI into neuromodulation decision workflows. An untrained GPT-4 model demonstrated high specificity but limited sensitivity, while the rule-based e-Health tool achieved moderate concordance with experts. A combined approach—merging structured algorithms for guideline adherence with adaptable language models for contextual reasoning—may represent the most realistic near-term pathway toward safe, explainable AI-assisted patient selection in SCS.

In the present study, the untrained LLM functioned effectively as a conservative filter, identifying clear exclusions with high specificity. However, its stringency would have excluded a substantial minority of patients who ultimately benefited from SCS. Conversely, the rule-based e-Health algorithm aligned more closely with expert judgment but lacked the nuance to interpret mixed pain presentations or psychosocial subtleties. Taken together, these findings support a layered decision-making strategy in which automated tools expedite straightforward cases, while the MDT remains essential for borderline or complex scenarios where prior surgical history, psychosocial resilience, or anatomical constraints critically influence outcomes.

Future work should focus on training decision-support models using granular, outcome-linked SCS data, elucidating mechanistic links between pain phenotype and waveform selection, and extending follow-up beyond two years to assess long-term durability and device optimization. By integrating these advances, centers can refine patient selection, improve efficiency, and most importantly, maximize the proportion of individuals who achieve meaningful and sustained pain relief through spinal cord stimulation.

Abbreviations

AI, artificial intelligence; DN4, Douleur Neuropathique 4; EQ-5D, EuroQol-5 Dimension; LLM, large language models; MDT, multidisciplinary team; SCS, spinal cord stimulator; VAS, visual analogue scale.

Data Availability Statement

The data presented in this study are available on request from the corresponding author due to privacy and ethical restrictions.

Ethics Statement

Institutional Review Board approval was obtained from *Comitato Etico Locale Palermo 1* (protocol #23; September 19, 2024). The ethical review board waived the requirement for informed consent due to the retrospective nature of the study and the use of fully anonymized data, as specified in the Methods section. All patient data were fully anonymized before analysis, and no identifiable information was shared with external servers. GPT-4 evaluations were performed in an offline, locally controlled environment to ensure compliance with the General Data Protection Regulation (GDPR) and institutional data-protection policies. The study conformed to the principles of the Declaration of Helsinki and current European Union ethical standards for medical AI research.

Funding

The authors received no departmental, institutional, governmental, or foundational support for this study.

Disclosure

GLB is a speaker for Abbott. AT and CLR are advisors for Augmend Health. FPD, SD, and PC declare no competing interests. LK receives grants and/or personal fees from Nevro, Medtronic, Neuralace, and Saluda outside of this submitted work. PS receives grants and/or personal fees from Nalu, Saluda, and Biotronik outside of this submitted work, also PS has a patent High dose Capsaicin with royalties paid to Averitas, a patent Vagus Nerve Stimulation licensed to electroCore, a patent Multiple books with royalties paid to Peter S Staats. SL receives grants and/or personal fees Abbott, Biotronik, Nalu, NeuroOne, Saluda, and Neuralace. TD receives grants and/or personal fees Abbott, Saluda, Biotronik, Saluda, and Boston Scientific; also TD has a patent Abbott pending. The authors report no other conflicts of interest in this work.

References

1. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. *Radiology*. 2023;307(2). doi:10.1148/RADIOL.230163
2. Al Kuwaiti A, Nazer K, Al-Reedy A, et al. A review of the role of artificial intelligence in healthcare. *J Pers Med*. 2023;13(6):951. doi:10.3390/JPM13060951
3. 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. doi:10.3389/FRAI.2023.1169595/PDF
4. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the Chat-GPT model. *Res Sq*. 2023. doi:10.21203/RS.3.RS-2566942/V1
5. Slitzky M, Yong RJ, Bianco G, Lo Emerick T, Schatman ME, Robinson CL. The future of pain medicine: emerging technologies, treatments, and education. *J Pain Res*. 2024;17:2833–2836. doi:10.2147/JPR.S490581
6. Lo Bianco G, Robinson CL, D'Angelo FP, et al. Effectiveness of generative artificial intelligence-driven responses to patient concerns in long-term opioid therapy: cross-model assessment. *Biomedicines*. 2025;13(3):636. doi:10.3390/BIOMEDICINES13030636
7. Chlorogiannis DD, Apostolos A, Chlorogiannis A, et al. The role of ChatGPT in the advancement of diagnosis, management, and prognosis of cardiovascular and cerebrovascular disease. *Healthcare*. 2023;11(21):2906. doi:10.3390/HEALTHCARE11212906
8. Kassab J, Nasr L, Gebrael G, et al. AI-based online chat and the future of oncology care: a promising technology or a solution in search of a problem? *Front Oncol*. 2023:13. doi:10.3389/FONC.2023.1176617/PDF
9. Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: a systematic review. *Health Promot Perspect*. 2023;13(3):183–191. doi:10.34172/HPP.2023.22
10. Kao HJ, Chien TW, Wang WC, Chou W, Chow JC. Assessing ChatGPT's capacity for clinical decision support in pediatrics: a comparative study with pediatricians using KIDMAP of Rasch analysis. *Medicine*. 2023;102(25):E34068. doi:10.1097/MD.0000000000034068
11. Stoneham S, Livesey A, Cooper H, Mitchell C. ChatGPT versus clinician: challenging the diagnostic capabilities of artificial intelligence in dermatology. *Clin Exp Dermatol*. 2024;49(7):707–710. doi:10.1093/CED/LLAD402

12. Lo Bianco G, Cascella M, Li S, et al. Reliability, accuracy, and comprehensibility of ai-based responses to common patient questions regarding spinal cord stimulation. *J Clin Med*. 2025;14(5):1453. doi:10.3390/JCM14051453
13. Giorgino R, Alessandri-Bonetti M, Luca A, et al. ChatGPT in orthopedics: a narrative review exploring the potential of artificial intelligence in orthopedic practice. *Front Surg*. 2023:10. doi:10.3389/FSURG.2023.1284015/PDF
14. Lo Bianco G, Al-Kaisy A, Natoli S, et al. Neuromodulation in chronic pain management: addressing persistent doubts in spinal cord stimulation. *J Anesth Analg Crit Care*. 2025;5(1). doi:10.1186/S44158-024-00219-6
15. Cascella M, Shariff MN, Lo Bianco G, et al. Employing the artificial intelligence object detection tool YOLOv8 for real-time pain detection: a feasibility study. *J Pain Res*. 2024;17:3681–3696. doi:10.2147/JPR.S491574
16. Thomson S, Huygen F, Prangnell S, et al. Appropriate referral and selection of patients with chronic pain for spinal cord stimulation: european consensus recommendations and e-health tool. *Eur J Pain*. 2020;24(6):1169–1181. doi:10.1002/EJP.1562
17. Cruccu G, Garcia-Larrea L, Hansson P, et al. EAN guidelines on central neurostimulation therapy in chronic pain conditions. *Eur J Neurol*. 2016;23 (10):1489–1499. doi:10.1111/ENE.13103
18. Dworkin RH, O'Connor AB, Kent J, et al. Interventional management of neuropathic pain: neuPSIG recommendations. *Pain*. 2013;154 (11):2249–2261. doi:10.1016/J.PAIN.2013.06.004
19. Overview | spinal cord stimulation for chronic pain of neuropathic or ischaemic origin | guidance | NICE. Available from: https://www.nice.org.uk/guidance/ta159. Accessed July 1, 2025.
20. Lo Bianco G, Tinnirello A, Papa A, et al. Interventional pain procedures: a narrative review focusing on safety and complications. part 1 injections for spinal pain. *J Pain Res*. 2023;16:1637–1646. doi:10.2147/JPR.S402798
21. Palmer N, Guan Z, Chai NC. Spinal cord stimulation for failed back surgery syndrome – patient selection considerations. *Transl Perioper Pain Med*. 2019;6(3):81.
22. Karri J, Joshi M, Polson G, et al. Spinal cord stimulation for chronic pain syndromes: a review of considerations in practice management. *Pain Physician*. 2020;23(6):599–616.
23. Celestin J, Edwards RR, Jamison RN. Pretreatment psychosocial variables as predictors of outcomes following lumbar surgery and spinal cord stimulation: a systematic review and literature synthesis. *Pain Med*. 2009;10(4):639–653. doi:10.1111/J.1526-4637.2009.00632.X
24. Sparkes E, Duarte RV, Mann S, Lawrenc TR, Raphael JH. Analysis of psychological characteristics impacting spinal cord stimulation treatment outcomes: a prospective assessment. *Pain Physician*. 2015;18(3):E369–E378. doi:10.36076/ppj.2015/18/e369
25. Thomson S, Huygen F, Prangnell S, et al. Applicability and validity of an e-health tool for the appropriate referral and selection of patients with chronic pain for spinal cord stimulation: results from a european retrospective study. *Neuromodulation*. 2023;26(1):164–171. doi:10.1016/j.neurom.2021.12.006
26. irr: various coefficients of interrater reliability and agreement. *CRAN: contributed Packages*. 2005. doi:10.32614/CRAN.PACKAGE.IRR.
27. Altman DG. *Practical Statistics for Medical Research*. 1990:404–409. doi:10.1201/9780429258589
28. Lee EJ, Edgerton ML, Buccilli B, Telkes I, Harland T, Pilitsis JG. Prediction of response to spinal cord stimulation using machine learning based on radiomics and patient-reported outcomes. *Neurosurgery*. 2025. doi:10.1227/NEU.0000000000003715