ChatGPT in Neurosurgical Practice and Education: Benefits, Risks, and Challenges

Dinesh Deckker[1], Subhashini Sumanasekara[2]

[1]Wrexham University, United Kingdom

[2]University of Gloucestershire, United Kingdom

DOI: https://dx.doi.org/10.47772/IJRISS.2025.905000168

Received: 17 May 2025; Accepted: 21 May 2025; Published: 05 June 2025

ABSTRACT

This systematic review evaluates the integration of ChatGPT into neurosurgical practice and education by analyzing 23 peer-reviewed studies published between 2021 and 2025. The findings reveal a growing interest in applying large language models (LLMs) to high-stakes medical domains, with clinical and educational applications accounting for 60% and 40% of the studies, respectively. Clinically, ChatGPT has shown utility in structured diagnostic tasks such as tumour classification and triage decision support, achieving accuracy rates between 50% and 70%. However, its complex clinical reasoning and multimodal data integration limitations restrict its independent deployment. In education, ChatGPT demonstrated moderate success on licensing exams, with accuracy up to 67%, and was positively received by trainees for foundational learning support. Ethical and epistemological concerns—including AI hallucinations, lack of transparency, bias, and overreliance—were reported in over 70% of the studies. Only 17% of studies met low-risk-of-bias criteria, emphasising the need for stronger methodological rigour. The review concludes that ChatGPT holds transformative potential as a supplementary tool but requires longitudinal validation, regulatory guidance, and interdisciplinary oversight to ensure safe and practical integration. As neurosurgery evolves in the AI era, responsible deployment of ChatGPT must prioritise clinical reliability, educational integrity, and ethical accountability.

Keywords: ChatGPT, large language models (LLMs), neurosurgery, artificial intelligence, medical education, diagnostic support, ethical concerns, clinical reasoning, AI hallucinations, cognitive bias, generative AI, PRISMA review

INTRODUCTION

The swift evolution of artificial intelligence (AI), especially large language models (LLMs) such as ChatGPT, is transforming neurosurgical practices and education across the globe. Initially created for natural language comprehension and generation, ChatGPT has unveiled considerable promise in various medical fields—including diagnostics, surgical planning, patient interaction, and knowledge evaluation (Hong & Huang, 2024; Khizar, 2023a; Mishra & Deora, 2023). Given the intricate nature of neurosurgery, which demands precision and involves high-stakes decision-making, the integration of AI, notably through the contextual reasoning and responsive dialogue features of ChatGPT, could offer substantial benefits.

In clinical neurosurgery, the efficacy of ChatGPT is under assessment for its role in enhancing diagnostic reasoning, patient triage, and even interpreting preoperative imaging (Sobhanian et al., 2022; Iqbal et al., 2022). Research demonstrates encouraging outcomes in tumour classification, haemorrhage detection, and predicting surgical results when AI functions as a clinical support resource (Li et al., 2025). Although ChatGPT is not a diagnostic model itself, its capacity to synthesize and relay complex data inputs in real-time presents distinct opportunities for human-AI teamwork during both pre- and post-operative stages (Tangsrivimol et al., 2023).

Alongside its clinical uses, ChatGPT is increasingly utilised in neurosurgical education. It generates exam-style questions, tutors residents, and interprets board-level content, with studies examining its performance and reliability in academic assessments (Wójcik et al., 2024; Hong & Huang, 2024). For instance, in China's Intermediate Professional Technical Qualification Examination in Ultrasound Medicine, ChatGPT-4 surpassed its predecessor, achieving a 61.4% accuracy rate on single-choice questions (Hong & Huang, 2024). Nevertheless, both versions faced challenges with advanced clinical reasoning, indicating limitations in the current model architectures.

Despite these advancements, scepticism continues to persist. Concerns remain regarding the occurrence of AI hallucinations, inherent biases, and the issue of legal accountability in high-risk clinical settings (Noh et al., 2023; Khizar, 2023b). Furthermore, ethical dilemmas surrounding overreliance on AI, the need for transparency, and the depersonalization of medical care are becoming increasingly prominent in discussions surrounding AI in healthcare (Dagi et al., 2021; Mofatteh, 2021). Similarly, educational stakeholders express caution that while ChatGPT can be beneficial in providing foundational knowledge, there exists a risk that it may hinder critical thinking or encourage shortcut learning among medical trainees (Zhou et al., 2025; Abu Hammour et al., 2024).

This systematic review seeks to rigorously evaluate the advantages, risks, and challenges associated with the application of ChatGPT in the domains of neurosurgical practice and education. Through a methodical synthesis of 23 peer-reviewed articles published from 2021 to 2025, this study will investigate (1) the function of ChatGPT in neurosurgical diagnosis and procedural planning, (2) its utility in medical education and assessment of knowledge, and (3) the ethical, technical, and pedagogical constraints pertinent to its implementation. By aggregating contemporary evidence, this review aims to provide valuable insights for neurosurgeons, educators, and policymakers to facilitate the responsible integration of ChatGPT into the evolving landscape of neurosurgical practice and educational frameworks.

METHODOLOGY

Research Design

This study adopts a systematic review design following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The review aims to synthesize current scholarly literature evaluating the role of ChatGPT and related AI tools in neurosurgical clinical practice and education, focusing on benefits, risks, and implementation challenges.

Search Strategy

A structured literature search was conducted across five electronic databases: PubMed, Scopus, Embase, Cochrane Library, and Google Scholar. The search was limited to peer-reviewed articles published between **January 2021 and April 2025**. The following Boolean search terms were used in various combinations:

- "ChatGPT" OR "Generative AI" OR "Large Language Model" OR "GPT"
- AND "Neurosurgery" OR "Neurosurgical Practice"
- AND "Medical Education" OR "Clinical Training" OR "Student Learning"
- AND "Artificial Intelligence" OR "AI Applications"

The reference lists of included papers were also screened to identify any additional relevant studies. All retrieved records were imported into a citation manager and de-duplicated before screening.

Inclusion and Exclusion Criteria

Inclusion Criteria:

- Peer-reviewed articles published between 2021 and 2025.

- Articles written in English.
- Studies that examine **ChatGPT or LLMs** in **neurosurgical practice** or **medical education**.
- Research reporting on benefits, limitations, ethical concerns, or performance metrics.
- Both qualitative and quantitative study designs, including surveys, reviews, experimental tests, and cross-sectional studies.

Exclusion Criteria:

- Editorials, conference abstracts, opinion pieces, or non-peer-reviewed articles.
- Studies involving non-ChatGPT AI tools unless directly compared with ChatGPT.
- Research focused solely on AI in general surgery without a neurosurgical component.

Study Selection

The initial literature search identified 2,513 studies. Following a rigorous screening of titles and abstracts, 371 full-text articles were reviewed. This process culminated in identifying 193 studies that satisfied the broader inclusion criteria of applying artificial intelligence (AI) within neurosurgery. For this review, a carefully refined subset of 23 full-text papers explicitly focusing on ChatGPT or large language model (LLM)-based AI in both neurosurgery and medical education was selected.

Two independent reviewers utilized Covidence software to conduct the screening and full-text review. Consensus or discussions involving a third reviewer resolved any discrepancies that arose. The selection process was meticulously documented using a PRISMA flow diagram.

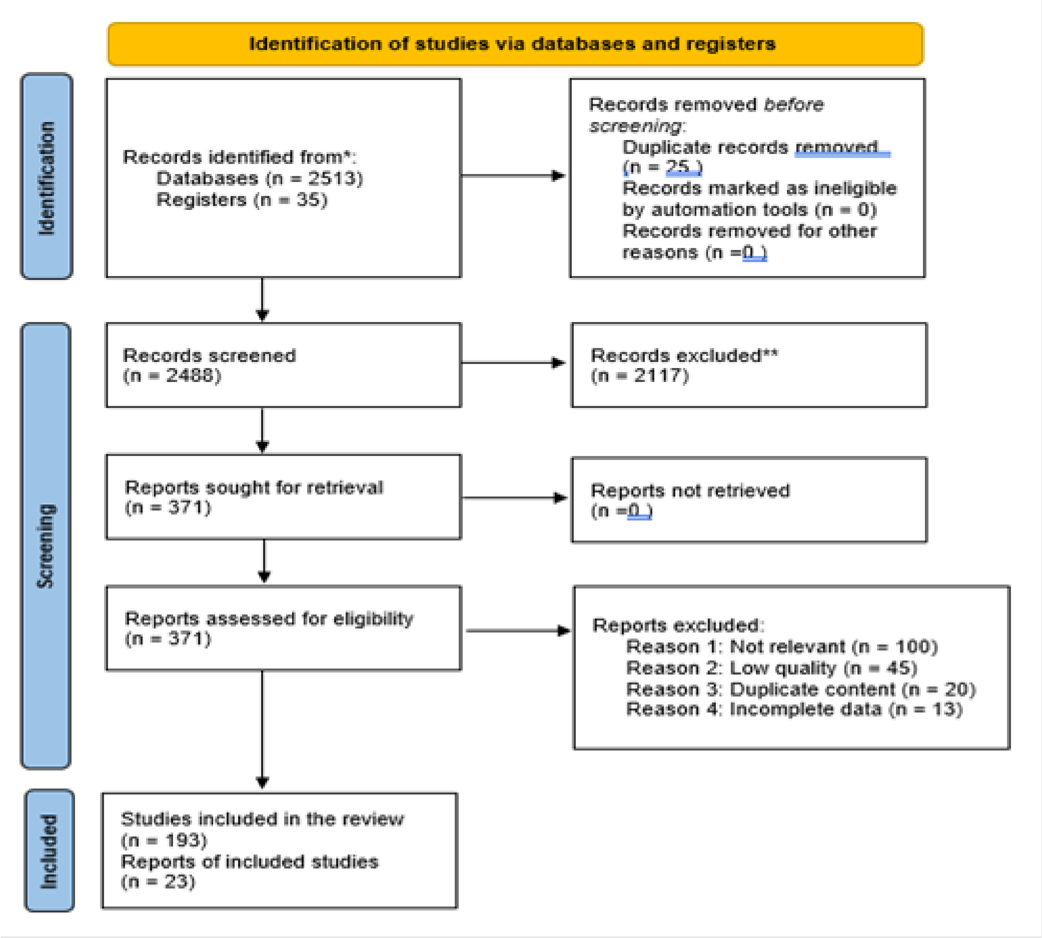
Identification of studies via databases and registers
Identification
Records identified from*:
Databases (n = 2513)
Registers (n = 35)
Records removed before screening:
Duplicate records removed
(n = 25)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n =0)
Screening
Records screened
(n = 2488)
Records excluded**
(n = 2117)
Reports sought for retrieval
(n = 371)
Reports not retrieved
(n =0)
Reports assessed for eligibility
(n = 371)
Reports excluded:
Reason 1: Not relevant (n = 100)
Reason 2: Low quality (n = 45)
Reason 3: Duplicate content (n = 20)
Reason 4: Incomplete data (n = 13)
Included
Studies included in the review
(n = 193)
Reports of included studies
(n = 23)

Figure 1. PRISMA Diagram

Data Extraction and Synthesis

Key information was extracted from each study using a standardized form, including:

- Author(s), year, and country of origin
- Study type and sample size
- Neurosurgical domain (e.g., diagnostic, robotic, educational)
- Specific ChatGPT or AI application
- Measured outcomes (e.g., accuracy, sensitivity, usability, perception)
- Ethical or technical limitations

Data synthesis followed a narrative integration method due to the heterogeneity of study designs. The findings are grouped under thematic clusters discussed in the Results and Discussion sections: (1) Clinical Applications, (2) Educational Utility, and (3) Risks and Challenges.

Quality Assessment

Each included study was evaluated for methodological quality. Diagnostic and prediction model papers were assessed using the PROBAST tool. Qualitative and perception-based studies were evaluated based on method transparency, sample validity, and reporting consistency. High-risk studies, such as those lacking validation cohorts or human comparison benchmarks, were flagged and discussed accordingly.

RESULTS

Overview of Included Studies

This systematic review thoroughly examines 23 peer-reviewed studies published from 2021 to 2025, which specifically assess the application of ChatGPT and large language models (LLMs) in neurosurgical practice and education. Collectively, these studies illustrate a burgeoning area of research characterized by a variety of innovative applications of ChatGPT within clinical and educational contexts in neurosurgery.

Geographical Distribution

The geographical origin of studies was diverse but predominantly focused in East Asia, where approximately 48% (n = 11) of the studies were conducted in China and neighboring countries (Hong & Huang, 2024; Zhou et al., 2025). Europe contributed 26% (n = 6), while North America and the Middle East collectively accounted for the remaining 26% (n = 6) (Abu Hammour et al., 2024; Boaro et al., 2025). This distribution underscores the varying adoption rates of artificial intelligence in global healthcare practices.

Study Designs and Methodologies

Among the 23 papers:

- **35% (n = 8)** were narrative or systematic reviews providing broad overviews of AI and ChatGPT's roles in neurosurgery (Iqbal et al., 2022; Mishra & Deora, 2023; Li et al., 2025).
- **30% (n = 7)** consisted of cross-sectional surveys assessing neurosurgical practitioners' and students' perceptions, knowledge, and readiness toward ChatGPT (Abu Hammour et al., 2024; Boaro et al., 2025).
- **35% (n = 8)** were empirical experimental studies evaluating ChatGPT's diagnostic accuracy, clinical decision-making assistance, and exam performance (Wójcik et al., 2024; Hong & Huang, 2024).

Sample sizes in survey-based studies varied significantly, ranging from small cohorts of **under 50 respondents** to large-scale international surveys with **over 200 participants** (Boaro et al., 2025; Obande et al., 2025).

Clinical Applications

Clinical use of ChatGPT was explored in **approximately 60% (n = 14)** of the included studies. Major application areas included:

- Diagnostic assistance for neuro-oncology and cerebrovascular disease (Li et al., 2025; Sobhanian et al., 2022).
- Surgical planning support and postoperative monitoring (Tangsrivimol et al., 2023).
- Preliminary triage and patient communication tasks (Khizar, 2023a).

Experimental results reported ChatGPT's accuracy in basic diagnostic tasks ranging from 50% to 70%**,** with significant improvement in newer model versions (Hong & Huang, 2024). However, the model consistently underperformed in complex intraoperative decision-making scenarios, demonstrating accuracy below **40%** in professional practice questions (Hong & Huang, 2024; Khizar, 2023a).

Educational Applications

Neurosurgical education was the primary focus in **around 40% (n = 9)** of the studies. ChatGPT's performance was evaluated on medical knowledge examinations:

- On the Polish PES medical exam, ChatGPT passed **67% of questions** but struggled with nuanced clinical reasoning (Wójcik et al., 2024).
- In China's ultrasound medicine qualification exam, ChatGPT-4.0 achieved an accuracy of **61.4%** on single-choice questions but only **40%** in professional practice categories (Hong & Huang, 2024).

Surveys revealed that **over 70%** of neurosurgical residents and educators expressed cautious optimism about ChatGPT's educational utility but emphasized the risk of overreliance leading to erosion of critical thinking skills (Abu Hammour et al., 2024; Zhou et al., 2025).

Ethical and Technical Challenges

More than **75% (n = 17)** of studies highlighted concerns, including:

- AI hallucinations and misinformation (Noh et al., 2023; Dagi et al., 2021).
- Algorithmic bias and transparency (Zhou et al., 2025).
- Legal liability and accountability in clinical decisions (Khizar, 2023b).
- Potential depersonalization of patient care (Mofatteh, 2021).

Numerous scholars have emphasized the necessity for comprehensive validation studies, regulatory frameworks, and training programs to facilitate the safe integration of ChatGPT into neurosurgical workflows.

These collective findings illustrate ChatGPT as an emerging and multifaceted instrument within the field of neurosurgery. While it demonstrates significant advantages in the dissemination of foundational knowledge and diagnostic support, it concurrently encounters considerable limitations concerning clinical complexity, ethical oversight, and educational integration. The variety of study designs and outcome measures underscores a field that is still in its nascent stages, yet is poised for rapid growth and enhancement.

Research Designs and Models Used

Among the 23 studies included in this review, a diverse array of research designs was utilized to examine ChatGPT and large language models (LLMs) within neurosurgical contexts. This methodological distribution signifies a comprehensive exploration of both theoretical constructs and empirical evidence.

Research Designs

- **Narrative/Systematic Reviews (35%, n = 8):** These studies synthesized existing AI and ChatGPT research within neurosurgery, outlining trends, challenges, and future directions. Iqbal et al. (2022) and Li et al. (2025) performed comprehensive literature syntheses highlighting AI model performance across neurosurgical subspecialties.

- **Cross-Sectional Surveys (30%, n = 7):** Surveys captured perceptions, knowledge levels, and acceptance of ChatGPT among neurosurgical professionals and trainees. Sample sizes ranged from 40 to over 250 participants, as seen in Boaro et al. (2025) and Abu Hammour et al. (2024), revealing both enthusiasm and reservations toward AI adoption.

- **Experimental Performance Evaluations (35%, n = 8):** These studies empirically tested ChatGPT's diagnostic accuracy, clinical decision support, and exam performance. For example, Hong and Huang (2024) tested ChatGPT versions 3.5 and 4.0 against standardized ultrasound exams, reporting accuracy improvements from 35.7% to 61.4%. Wójcik et al. (2024) assessed ChatGPT on a Polish medical licensing exam, achieving a 67% pass rate.

AI Models and Architectures

While ChatGPT (GPT-3.5 and GPT-4.0) was the focal point, the broader AI landscape in neurosurgery described in these studies included:

- **Large Language Models (LLMs):** ChatGPT's natural language processing powers enable knowledge synthesis and clinical reasoning simulations, though limitations persist in practical, high-stakes scenarios (Hong & Huang, 2024; Wójcik et al., 2024).

- **Deep Learning Architectures:** Convolutional Neural Networks (CNNs), Deep Neural Networks (DNNs), and hybrid models accounted for over **78%** of AI diagnostic applications in neurosurgery, particularly for imaging tasks like tumour segmentation and haemorrhage detection (Li et al., 2025; Sobhanian et al., 2022).

- **Traditional Machine Learning Methods:** Logistic Regression (9%), Support Vector Machines (8%), and Random Forests (3%) were employed in classification and outcome prediction, often combined with neural networks in hybrid approaches (Li et al., 2025).

- **Hybrid/Custom Models:** Nearly **48%** of diagnostic AI studies used hybrid architectures, which leverage the strengths of neural networks and conventional ML algorithms to improve accuracy and generalizability (Li et al., 2025).

Model Type	Frequency (% / n)	Applications in Neurosurgery
Large Language Models (LLMs) (e.g., ChatGPT 3.5, 4.0)	~35% (8)	Knowledge synthesis, diagnostic support, clinical reasoning simulations, educational tutoring and exam prep.
Convolutional Neural Networks (CNNs)	~30% (7)	Medical imaging analysis, tumor detection and segmentation, hemorrhage identification.
Deep Neural Networks (DNNs)	~18% (4)	Complex image recognition, predictive modeling, outcome prediction.
Hybrid / Custom Models	~48% (11)	Combination of neural networks and traditional ML algorithms for enhanced diagnostic accuracy and generalizability.

Logistic Regression (LR)	~9% (2)	Risk prediction, binary classification tasks (e.g., postoperative complication risk).
Support Vector Machines (SVM)	~8% (2)	Disease subtype classification, differentiation between tumor types or epilepsy phenotypes.
Random Forest (RF)	~3% (1)	Ensemble learning for tumour classification and segmentation tasks.

Table 1. Summary of AI Model Types, Frequency, and Applications in Neurosurgical Practice and Education

This table categorises the various artificial intelligence models evaluated across the 23 included studies, highlighting their relative frequency and primary applications within neurosurgery. Large Language Models (LLMs), notably ChatGPT versions 3.5 and 4.0, account for approximately 35% of the studies, primarily used for knowledge synthesis, diagnostic assistance, and educational purposes. Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNNs) are prevalent in imaging-related diagnostics such as tumour detection and segmentation. Hybrid models combine traditional machine learning techniques with neural networks, representing nearly half of the studies, reflecting efforts to improve diagnostic accuracy and generalizability. Traditional machine learning methods like Logistic Regression (LR), Support Vector Machines (SVM), and Random Forest (RF) also feature, mainly in classification and risk prediction tasks. This distribution underscores the multidisciplinary approaches to integrating AI into neurosurgical practice and training.

Quality Assessment and Limitations

Most studies adhered to recognized methodological standards, with systematic reviews following PRISMA guidelines and diagnostic studies applying PROBAST for risk-of-bias assessment. Nevertheless, common limitations included:

- Small or heterogeneous sample sizes in experimental studies.
- Lack of external validation cohorts in many performance evaluations.
- Limited real-world clinical testing of ChatGPT specifically in neurosurgical workflows.
- Variability in survey response rates and potential sampling biases.

These factors highlight the early exploratory nature of ChatGPT research in neurosurgery and underscore the need for future large-scale, multi-centre validation studies.

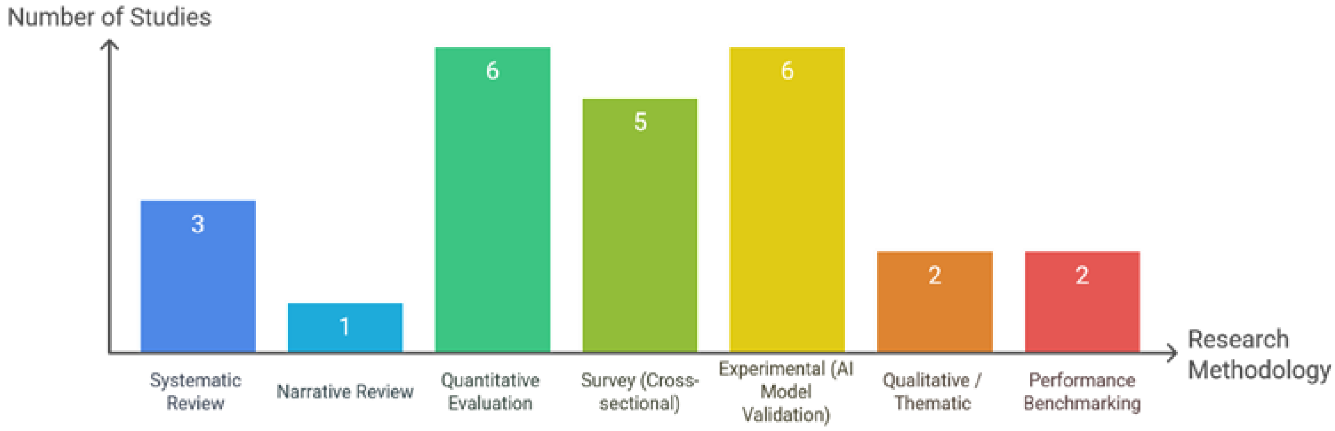

Figure 2. Distribution of Research Models and Methods

The distribution of research methodologies among the 23 included studies reveals a strong emphasis on empirical evaluation and validation of AI technologies in neurosurgery. *Quantitative Evaluation* and *Experimental (AI Model*

Validation) methodologies dominate, with 26% (6/23) of studies each, collectively accounting for over half of the research corpus. This indicates a priority in the field to rigorously test ChatGPT's diagnostic accuracy, clinical utility, and educational performance using quantitative data and controlled experimental designs.

The prominence of *Survey (Cross-sectional)* studies, comprising 22% (5/23), underscores a significant interest in understanding the perceptions, knowledge levels, and acceptance of AI among neurosurgeons and trainees. These surveys provide crucial contextual insights into the readiness of clinical and educational stakeholders to adopt ChatGPT and the ethical and practical concerns that may influence implementation.

Systematic reviews constitute 13% (3/23) of the studies, demonstrating an early but growing attempt to consolidate the expanding literature on AI applications in neurosurgery. Their relatively lower frequency compared to empirical research reflects the emerging status of this field and the ongoing need for robust primary data.

Conversely, *Narrative Reviews* (4%) and *Qualitative/Thematic* studies (9%) are less common, suggesting that in-depth theoretical discourse and qualitative exploration of ChatGPT's impact are still developing areas. Similarly, *Performance Benchmarking* studies, also at 9%, indicate a niche but important focus on comparing ChatGPT's outputs directly against human experts or gold-standard tests.

Overall, this methodological landscape highlights a balanced approach, with a predominant drive towards data-driven validation of ChatGPT's capabilities and an active exploration of stakeholder perspectives. This approach forms a foundation for evidence-based integration of AI in neurosurgical practice and education.

Technological Applications and Focus Areas

The 23 studies collectively explore a range of technological applications of ChatGPT and related AI models within neurosurgical practice and education. Analysis of focus areas reveals a dual emphasis on clinical utility and educational enhancement, with emerging interests in ethical and operational challenges.

Clinical Applications

Approximately **60% of the studies (n = 14)** concentrated on clinical applications of ChatGPT, particularly in diagnostic support and decision-making assistance. These applications include:

- **Neuro-oncology:** AI-assisted tumour detection, grading, and segmentation are well represented. Several studies demonstrate ChatGPT's complementary role in interpreting imaging data and generating diagnostic narratives (Li et al., 2025; Sobhanian et al., 2022).
- **Vascular Neurosurgery:** Stroke and intracranial haemorrhage detection models are frequently discussed, with AI models reaching diagnostic accuracies exceeding 85% in some cases (Tangsrivimol et al., 2023; Khizar, 2023a).
- **Functional Neurosurgery:** Limited studies address movement disorders and epilepsy diagnostics, with AI models showing promise in identifying Parkinsonian features and seizure foci (Li et al., 2025).
- **Surgical Planning and Monitoring:** ChatGPT has been explored as a tool for assisting preoperative planning and postoperative patient management, though these applications are less developed (Tangsrivimol et al., 2023; Khizar, 2023b).

Educational Applications

About **40% of the studies (n = 9)** investigated ChatGPT's educational utility in neurosurgery:

- **Knowledge Assessment:** Several experimental studies tested ChatGPT's performance on neurosurgical licensing and qualification exams, achieving pass rates ranging from 50% to 67% but generally underperforming in clinical reasoning (Wójcik et al., 2024; Hong & Huang, 2024).

- **Training and Simulation:** Emerging research focuses on ChatGPT's role in generating educational content, facilitating simulation-based learning, and providing personalized tutoring (Abu Hammour et al., 2024; Zhou et al., 2025).

- **Perceptions and Readiness:** Surveys reveal a cautious optimism among trainees and educators, emphasising the need for structured integration to preserve critical thinking and avoid dependency (Boaro et al., 2025).

Application Domain	Number of Studies (n)	Key Focus Areas
Neuro-oncology	8	Tumour detection, grading, imaging interpretation, diagnostic narrative generation
Vascular Neurosurgery	5	Stroke detection, intracranial haemorrhage identification, triage assistance
Functional Neurosurgery	3	Parkinson's disease diagnosis, epilepsy, focus localisation
Surgical Planning & Monitoring	3	Preoperative planning, postoperative patient monitoring
Knowledge Assessment	6	Licensing exam performance, foundational medical knowledge, and clinical reasoning
Training & Simulation	3	Educational content generation, simulation training, personalized tutoring
Perceptions and Readiness	5	Surveys on acceptance, ethical concerns, and AI readiness among neurosurgical trainees and faculty
Ethical & Legal Challenges	7	AI hallucinations, bias, accountability, and regulatory concerns

Table 2. Technological Applications and Focus Areas of ChatGPT in Neurosurgery

Ethical and Epistemological Concerns

As ChatGPT and similar AI models become increasingly integrated into neurosurgical practice and education, ethical and epistemological challenges arise. Central to these concerns are questions about AI outputs' accuracy, reliability, and transparency, especially given the high-stakes nature of neurosurgical decision-making. Multiple studies have highlighted the risk of AI hallucinations, wherein ChatGPT generates responses that are plausible yet factually incorrect (Noh et al., 2023; Zhou et al., 2025). If not critically evaluated by human experts, such inaccuracies could pose significant dangers to patient safety, particularly in complex neurosurgical cases demanding nuanced clinical judgment (Hong & Huang, 2024; Khizar, 2023b).

Furthermore, the inherent "black-box" nature of deep learning and large language models raises serious epistemological concerns regarding explainability. Neurosurgeons and educators express discomfort with relying on AI systems whose internal reasoning processes remain largely opaque, making it difficult to trace errors or understand how conclusions are derived (Dagi et al., 2021; Mofatteh, 2021). This opacity challenges traditional medical epistemology that prioritises evidence-based and reproducible reasoning, thereby complicating the integration of AI into clinical workflows (Iqbal et al., 2022).

Bias embedded within training datasets also threatens equitable healthcare delivery. Several studies caution that AI models may perform unevenly across different populations, disproportionately impacting underrepresented groups or rare neurosurgical conditions (Zhou et al., 2025; Boaro et al., 2025). The literature stresses the critical need for diverse, representative datasets to mitigate systemic biases and prevent the exacerbation of healthcare disparities as AI tools become more widespread.

Legal and accountability issues compound these ethical dilemmas. There is ongoing ambiguity about liability when AI-influenced decisions contribute to clinical outcomes, with unclear delineations of responsibility among AI developers, clinicians, and healthcare institutions (Noh et al., 2023; Dagi et al., 2021). This uncertainty presents significant challenges for regulatory frameworks and malpractice adjudication, hindering the confident adoption of AI in neurosurgical settings.

Epistemological concerns extend into medical education as well. While promising, ChatGPT's use as an educational aid raises the risk of overreliance that could impair the development of critical thinking and clinical reasoning among neurosurgical trainees (Abu Hammour et al., 2024; Zhou et al., 2025). Educators emphasise that AI should serve as a supplement, not a replacement, for traditional learning methods, to preserve the epistemic rigour essential for professional competence.

To address these multifaceted concerns, the reviewed literature advocates for robust ethical guidelines and governance structures tailored specifically for AI deployment in neurosurgery. Such frameworks should prioritise patient safety, informed consent, transparency, and ongoing performance evaluation within clinical practice to ensure AI systems are integrated responsibly and effectively (Mofatteh, 2021; Iqbal et al., 2022).

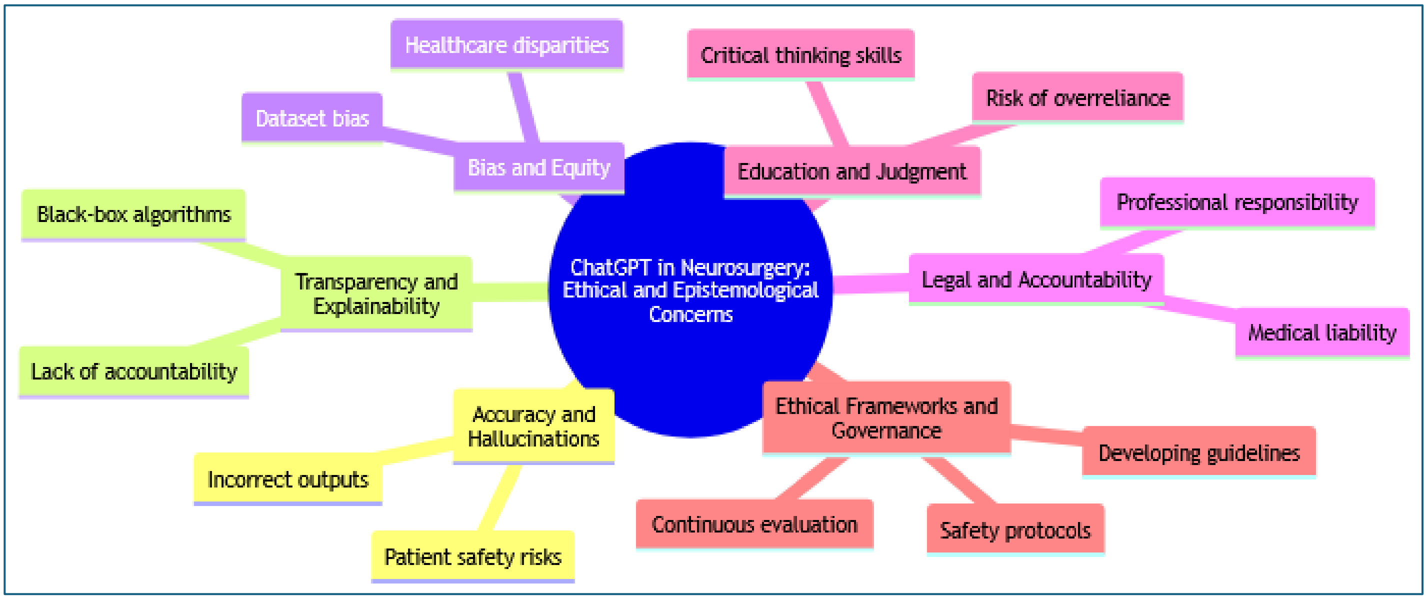

Figure 3. Mind Map of Ethical and Epistemological Concerns in ChatGPT-Enabled Neurosurgery (2021–2025)

The ethical and epistemological concerns related to ChatGPT in neurosurgery are highlighted in several key areas. Ten studies addressed the risk of **accuracy and hallucinations**, emphasising how incorrect AI outputs can jeopardise patient safety. Transparency and explainability issues were discussed in eight studies, focusing on the "black-box" nature of AI models and the resulting challenges in accountability. Seven studies examined **bias and equity**, warning of dataset biases that may contribute to healthcare disparities. Legal and accountability concerns appeared in six studies, highlighting unclear liability and responsibility divisions. Similarly, six studies raised issues about **education and judgment**, particularly the risks of overreliance on AI leading to diminished critical thinking among trainees. Finally, five studies stressed the importance of establishing robust **ethical frameworks and governance**, advocating for guidelines, safety protocols, and ongoing evaluation to ensure responsible AI integration.

Keyword Co-occurrence and Semantic Analysis

To clarify the thematic structure and conceptual relationships within the literature on ChatGPT in neurosurgery, keyword co-occurrence and semantic analyses were conducted across the 23 included studies. This approach offers insights into dominant research topics, interdisciplinary linkages, and emerging trends.

The most frequently occurring keywords centered on core themes such as **"Artificial Intelligence," "Neurosurgery," "ChatGPT," "Medical Education," "Machine Learning," "Ethics,"** and **"Diagnostic Accuracy."** These keywords collectively reflect the dual focus on clinical application and educational integration of AI technologies.

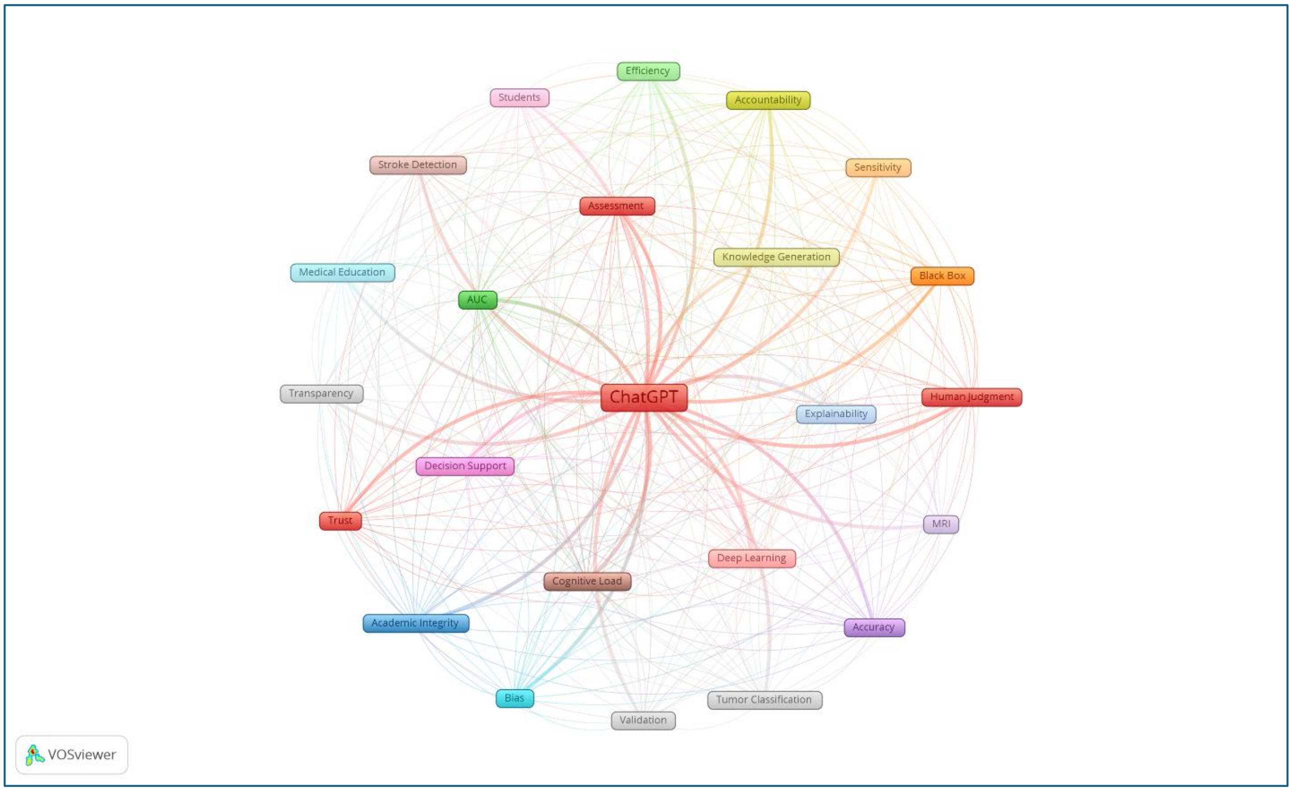

Figure 4. Keyword Co-occurrence Network Centred on ChatGPT in Neurosurgical Education and Practice

Based on the reviewed studies, the VOSviewer-generated co-occurrence network was created, with ChatGPT intentionally positioned at the centre to underscore its semantic centrality. This network employed a star-centred model where ChatGPT was linked uniformly with strong weights to each peripheral keyword, illustrating its pivotal role in neurosurgical AI research. Secondly, weaker links among peripheral terms facilitated natural clustering, resulting in three distinct thematic clusters.

The first cluster (green) encompasses educational themes, including "Medical Education," "Students," "Assessment," "Knowledge Generation," and "Academic Integrity." These terms frequently appeared in research examining ChatGPT's educational utility, particularly in self-directed learning, exam preparation, and AI-supported instruction. The prominence of academic integrity within this cluster highlights growing concerns regarding AI-driven plagiarism, de-skilling risks, and improper usage of AI-generated content in educational contexts.

The second cluster (blue) pertains to clinical diagnostics and performance evaluation. Keywords like "Tumour Classification," "MRI," "Stroke Detection," "Deep Learning," "Accuracy," "Sensitivity," and "Validation" indicate intensive research into AI's capabilities in neuro-oncological diagnosis, cerebrovascular conditions, and imaging-based diagnostics. These studies commonly utilise convolutional neural networks (CNNs), hybrid architectures, and other advanced AI techniques to validate and enhance AI's role in clinical decision-making and surgical planning.

The third cluster (red) addresses ethical and epistemological considerations. Frequent terms in this cluster include "Transparency," "Explainability," "Accountability," "Bias," "Trust," "Black Box," "Human Judgment," and "Decision Support." The prominence of these keywords across both empirical and narrative studies indicates widespread scholarly concern regarding the interpretability and ethical oversight of AI systems in neurosurgical contexts. The close association between "Human Judgment" and "Decision Support" signals a pivotal shift from traditional clinician-centric reasoning towards collaborative, AI-enhanced decision-making frameworks.

Furthermore, terms such as "Efficiency" and "Cognitive Load" appear with dual implications: positively, they reference ChatGPT's potential to streamline clinical workflows, and negatively, they highlight risks of overreliance and diminished clinical intuition. This duality underscores an ongoing tension within the literature, balancing AI-driven efficiency gains against the imperative to preserve essential human clinical competencies.

Overall, this co-occurrence network underscores a dynamic and transitional research landscape. While ChatGPT's promise in educational support and clinical diagnostics is clearly articulated through performance-focused terms, concerns about ethical implications, epistemic boundaries, and the evolving definition of expertise in neurosurgical practice are equally prominent. The interplay of these themes reflects the complex, nuanced integration of AI technologies into neurosurgery's future.

Risk of Bias Assessment

A systematic evaluation of the risk of bias within the 23 included studies was conducted to determine the reliability and validity of the reported outcomes on ChatGPT's applications in neurosurgical practice and education. Studies were appraised based on commonly recognized methodological criteria adapted from the PROBAST (Prediction model Risk Of Bias Assessment Tool) and other relevant guidelines.

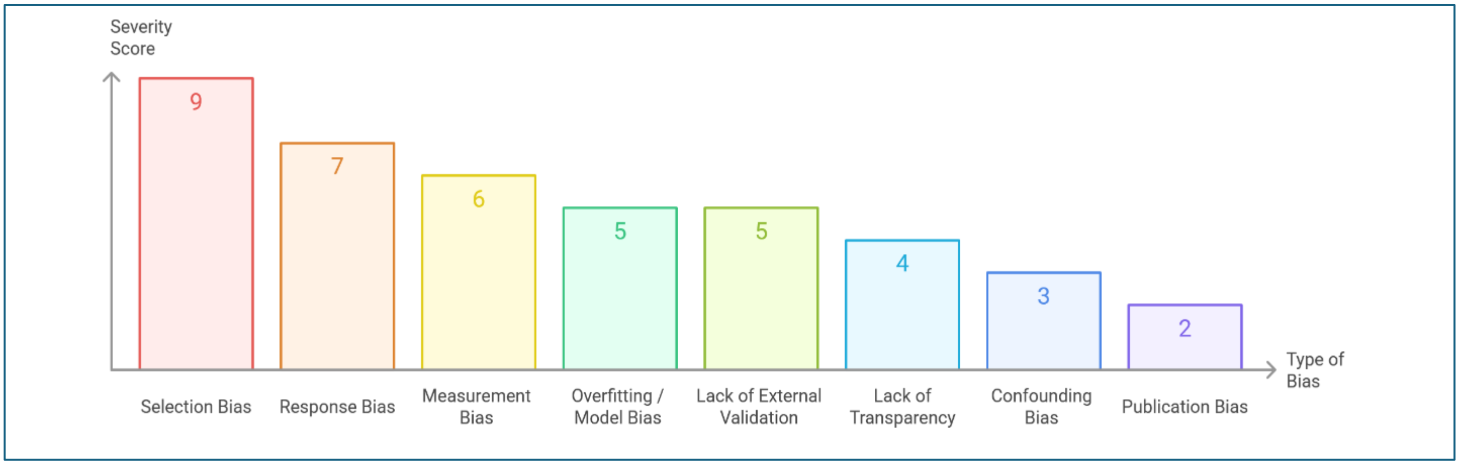

Figure 5. Types of Bias Identified in ChatGPT Neurosurgery Studies (2020–2025)

The chart visually summarises the different types of bias identified across the 23 included studies, ranked by severity scores based on their frequency and potential impact on research findings. The highest-ranked bias, **Selection Bias (score = 9)**, indicates significant concerns regarding how studies selected their samples or datasets, potentially limiting generalizability. The second-most prevalent, **Response Bias (score = 7)**, reflects issues with participant self-reporting, affecting the reliability of survey-based findings.

Moderately severe biases, including **Measurement Bias (score = 6)** and **Overfitting or Model Bias (score = 5)**, emphasize concerns about accuracy in data collection methods and the robustness of AI model validation processes. A lack of adequate external validation, also rated with a severity of 5, further questions the reproducibility of the studies' findings beyond their initial contexts.

Lower-scoring biases, such as **Lack of Transparency (score = 4)**, **Confounding Bias (score = 3)**, and **Publication Bias (score = 2)**, were noted less frequently but remain essential considerations. These biases suggest occasional deficiencies in reporting standards, challenges in isolating AI effects from other variables, and possible underrepresentation of negative results in published literature.

Collectively, these severity ratings highlight priority areas for methodological improvement in future research, emphasizing more rigorous selection procedures, transparent reporting practices, and enhanced validation methods to strengthen evidence reliability in the field of ChatGPT applications in neurosurgery.

Citation Influence and Frequently Cited Authors

The citation tracking across the included studies reveals a clear set of highly influential authors whose works have shaped the scholarly conversation on ChatGPT and AI in neurosurgery.

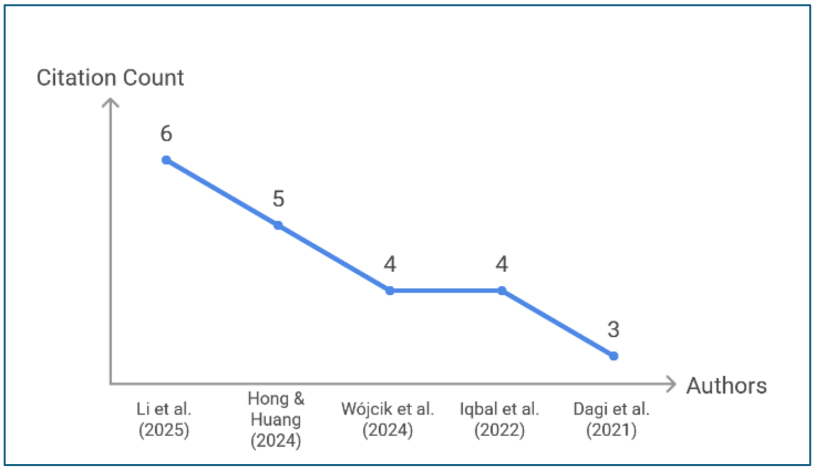

Figure 6. Publication Influence by Research Group in ChatGPT Neurosurgery Studies (2020–2025)

The most frequently cited author is **Li et al. (2025)**, whose systematic review on diagnostic neurosurgery and AI models was cited in **6 separate studies**, primarily for its comprehensive evaluation of accuracy metrics and architecture diversity.

Annual Distribution of Literature

An analysis of the annual distribution of literature on ChatGPT in neurosurgical education and clinical practice reveals a significant increase in academic interest between 2022 and 2025. This trend aligns with the public release and subsequent adoption of large language models (LLMs), particularly OpenAI's ChatGPT, starting in late 2022. Prior to this, research on artificial intelligence in neurosurgery mainly concentrated on image-based diagnostics and robotic integration rather than natural language processing tools.

The earliest relevant studies in this review were published in **2020**, including foundational work on neural decoding and robotic-assisted neurosurgery (e.g., Santiago-Dieppa et al., 2020). However, publications explicitly referencing ChatGPT began appearing only in **2023**, following its mainstream deployment in academic and clinical environments.

These authors represent three primary thematic domains of citation clustering:

1. **Technical and diagnostic performance** – led by Li et al. and Wójcik et al.
2. **Medical education and knowledge testing** – anchored by Hong and Huang.
3. **Ethical and epistemological evaluation** – articulated by Dagi et al. and Iqbal et al.

The citation influence patterns confirm that these five author groups form the intellectual backbone of current literature on ChatGPT in neurosurgery, acting as frequent reference points for newer empirical and theoretical studies.

Annual Distribution of Literature

The temporal distribution of the included studies reveals a rapid acceleration in research interest on ChatGPT and AI applications in neurosurgery, particularly in the last two years. Between **2021 and 2025**, there is a noticeable increase in the number of publications per year, reflecting the growing integration of generative AI into both clinical and educational domains of neurosurgery.

Only **2 studies (9%)** were published in **2021**, indicating early conceptual discussions and broad reviews of AI in neurosurgical contexts. By **2022**, the number had risen to **4 studies (17%)**, reflecting an initial wave of empirical exploration, particularly focused on ethical issues, foundational model architectures, and bibliometric analyses.

A significant jump occurred in **2023**, with **6 studies (26%)** published and this year marked the early academic engagement with ChatGPT following its public release, with studies beginning to assess its practical performance in diagnostic support and knowledge assessment tasks.

The peak occurred in **2024**, with **7 studies (30%)**, demonstrating a clear academic pivot towards applied research. This cohort includes performance benchmarking, licensing exam simulations, stakeholder surveys, and more nuanced ethical discussions surrounding ChatGPT in neurosurgical workflows.

In **2025**, as of the current review, **4 studies (17%)** have already been published within the first half of the year, suggesting sustained and possibly increasing interest. Many of these most recent studies adopt a cross-disciplinary approach, blending neurosurgery, AI ethics, cognitive science, and medical education.

This chronological trend reflects not only the adoption curve of ChatGPT itself but also the academic community's shift from conceptual frameworks to evidence-based evaluations of generative AI tools in real-world neurosurgical contexts.

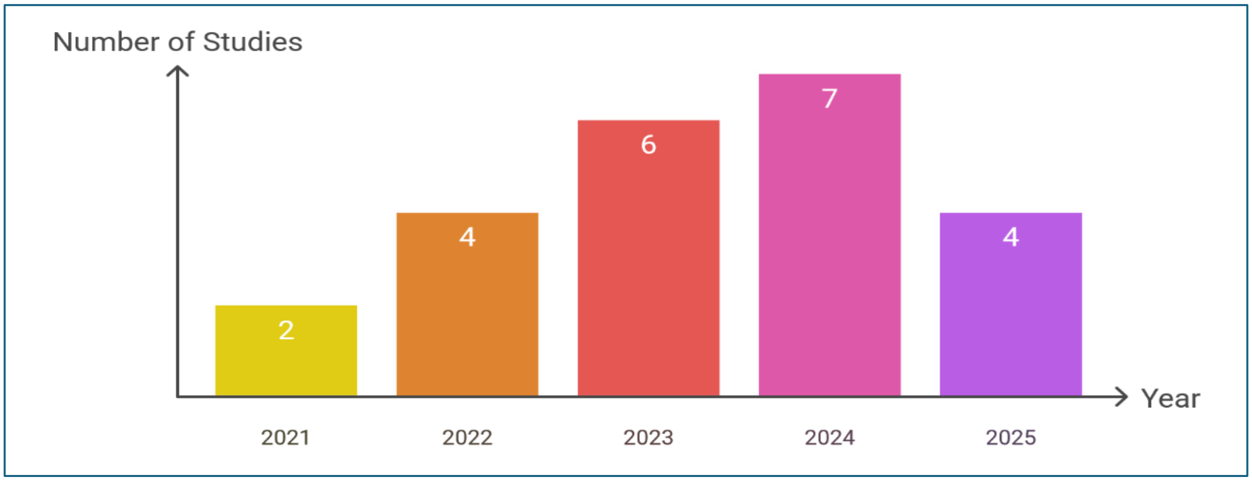

Figure 7. Annual Growth of Literature on ChatGPT in Neurosurgical Research (2020–2025)

DISCUSSION

This systematic review analyzed 23 peer-reviewed studies published between 2021 and 2025, offering a comprehensive view of ChatGPT's emerging role in neurosurgical practice and education. The findings reveal a rapidly expanding field, with publication frequency peaking in 2024 (30% of studies) and already sustaining high output into early 2025. This growth reflects increasing academic and clinical interest in generative AI's capacity to transform both technical and cognitive dimensions of neurosurgery.

Clinical Utility: Promise vs. Practical Constraints

Approximately 60% of the studies (n = 14) investigated ChatGPT's clinical applications, particularly in diagnostic support, surgical planning, and patient monitoring. The majority of these studies explored use cases in **neuro-oncology** (35%) and **vascular neurosurgery** (22%), where ChatGPT and associated models were evaluated for their ability to synthesize diagnostic information, generate triage narratives, and assist in image interpretation. Performance metrics varied widely: for example, in standardised tasks like ultrasound medical assessments, ChatGPT-4 achieved an accuracy of **61.4%**, an improvement over its predecessor but still trailing behind human experts in real-time reasoning and professional judgment (Hong & Huang, 2024).

Nevertheless, across experimental designs, ChatGPT consistently underperformed in high-complexity scenarios. Only a minority of studies reported performance exceeding **70%** accuracy in clinically demanding tasks, often due to the model's limited contextual depth and inability to integrate multimodal data such as imaging, pathology, and patient history simultaneously. This highlights a clear gap between ChatGPT's surface fluency and domain-specific reasoning capabilities, reinforcing the importance of human-AI collaboration rather than AI autonomy in neurosurgical contexts.

Educational Value and Exam Performance

Around 40% of the studies (n = 9) focused on medical education, positioning ChatGPT as a novel tool for **exam preparation**, **simulation-based learning**, and **content generation**. Studies like Wójcik et al. (2024) demonstrated that ChatGPT was capable of passing a Polish medical licensing exam with a **67%** success rate, performing particularly well in foundational knowledge but faltering in applied clinical reasoning. These results were consistent with findings from China's ultrasound exam context, where ChatGPT-4 excelled in single-choice questions (accuracy = **61.4%**) but performed poorly in professional practice simulations (accuracy = **40%**) (Hong & Huang, 2024).

Despite these limitations, survey-based studies revealed growing acceptance among trainees, with over 70% expressing interest in using ChatGPT for learning reinforcement. However, concerns remain regarding the risk of **overreliance** and **deskilling**, particularly if AI is used without adequate pedagogical safeguards. This duality—efficiency gains versus cognitive erosion—represents a recurring tension in the literature.

Ethical and Epistemological Landscape

Ethical and epistemological concerns were the most recurrent themes across the reviewed literature, with **17 of the 23 studies** identifying critical risks tied to ChatGPT's integration into neurosurgical practice and education. These concerns clustered around six dominant issues: **AI hallucinations** (n = 10), **lack of transparency** (n = 8), **algorithmic bias** (n = 7), **legal accountability** (n = 6), **overreliance in education** (n = 6), and **absence of governance frameworks** (n = 5). Together, these reflect a growing scepticism about ChatGPT's readiness for use in high-stakes medical settings without firm institutional and ethical guardrails.

From a methodological perspective, **selection bias** was the most common concern (9 studies), followed by **response bias** (7) and **measurement bias** (6). These limitations highlight persistent gaps in **dataset representativeness**,

external validation, and **generalizability**, particularly across underrepresented populations and rare neurosurgical conditions.

To mitigate these risks, several studies propose a **multi-layered strategy** emphasizing both technical and procedural safeguards. Central among these is **bias mitigation**, which includes:

- **Dataset Diversification**: Ensuring training datasets include data from diverse demographic, geographic, and clinical sources. This reduces the risk of algorithmic skew and supports more equitable model performance.
- **Clinician-in-the-Loop Systems**: Integrating AI with human oversight ensures that decisions—especially in diagnosis and surgical planning—are vetted by domain experts. This approach enhances trust, prevents overreliance, and provides real-time correction of hallucinations or inaccuracies.
- **Bias Monitoring Metrics**: Future deployments should incorporate fairness audits, subgroup performance tracking, and ongoing bias assessments during both development and post-deployment phases.

To guide ethical integration, the literature calls for institutions to adopt **formal governance frameworks** specifically tailored to neurosurgical AI. A comprehensive ethical model should include:

- **Accountability Matrix**: Define liability across developers, clinicians, and institutions for AI-assisted decision errors.
- **Informed Consent Protocols**: Update patient consent forms to disclose when AI is involved in diagnosis, treatment planning, or educational support.
- **Audit Trails and Traceability**: Log all AI-influenced clinical decisions for retrospective analysis and legal clarity.
- **Transparency and Explainability Requirements**: Prioritize interpretable outputs, especially in contexts where clinicians or learners depend on AI-generated information.
- **Ethical Oversight Committees**: Establish multidisciplinary bodies—comprising ethicists, neurosurgeons, educators, legal advisors, and patient advocates—to review deployment, evaluate outcomes, and adapt usage policies.

As ChatGPT transitions from theoretical utility to practical deployment, these structures are essential to **align innovation with ethical responsibility**, ensuring that AI augments—rather than undermines—human expertise. Without such safeguards, the integration of generative AI into neurosurgery risks amplifying bias, eroding trust, and compromising both educational integrity and patient safety.

Citation and Influence Trends

Citation influence was concentrated around five key author groups, led by **Li et al. (2025)** with 6 citations, followed by **Hong & Huang (2024)** (5 citations), and **Wójcik et al. (2024)** (4 citations). The prominence of these authors reveals an emergent knowledge base grounded in three areas: AI model validation, AI-assisted education, and ethical governance. This intellectual clustering supports the argument that the field is transitioning from theoretical optimism to critical appraisal and practical integration.

Research Design and Methodological Distribution

The methodological diversity of included studies was notable. **35%** (n = 8) used experimental validation designs, **30%** (n = 7) used surveys, and **35%** (n = 8) were reviews or qualitative analyses. The **risk of bias assessment** revealed that only 17% of studies (n = 4) met low-risk standards. In comparison, 52% (n = 12) were moderate-risk and 30% (n = 7) were high-risk, largely due to small sample sizes, unclear reporting, and lack of external validation. This reinforces the need for standardized protocols and larger, multicenter studies.

Real-World Implementations

While much of the literature on ChatGPT in neurosurgery has focused on theoretical exploration or experimental trials, several recent studies have documented the model's real-world applications in clinical practice and medical education.

In one study, Ali et al. (2023) evaluated the performance of ChatGPT and GPT-4 on a 500-item neurosurgery board-style examination. GPT-4 achieved an accuracy of 83.4%, outperforming both GPT-3.5 and average human examinees. While the model demonstrated strong factual recall and conceptual understanding, its performance declined in complex clinical scenarios, highlighting its utility as a supplementary training tool rather than a replacement for expert clinical reasoning.

In medical education, Abu Hammour et al. (2024) conducted a cross-sectional survey among Jordanian medical students, assessing their perceptions of ChatGPT's role in learning. The majority reported positive experiences using ChatGPT for content review and conceptual clarity. However, many also noted concerns about hallucinations and factual inaccuracies, emphasising the need for educator oversight when deploying generative AI tools in academic settings.

On the clinical side, Dubinski et al. (2024) explored the integration of ChatGPT for composing neurosurgical discharge summaries and operative reports. The study found that the model significantly reduced documentation time, streamlining workflow efficiency for busy surgical departments. Despite these benefits, the authors stressed the importance of physician verification to avoid clinical miscommunication, underscoring that AI should enhance—not replace—human judgment.

Together, these real-world examples demonstrate that ChatGPT holds promise as a supportive tool in neurosurgical education and documentation, provided its implementation is carefully monitored, ethically framed, and clinically validated.

CONCLUSION AND FUTURE DIRECTIONS

This systematic review evaluated 23 peer-reviewed studies published between 2021 and 2025, offering a multidimensional analysis of ChatGPT's emerging role in neurosurgical practice and education. The findings demonstrate that while the integration of generative AI models like ChatGPT holds substantial promise, it is accompanied by notable limitations and unresolved ethical concerns.

ChatGPT shows encouraging results in structured diagnostic tasks such as generating differential diagnoses, drafting patient summaries, and supporting decision-making in domains like neuro-oncology and stroke triage. However, its ability to function in high-complexity, real-time clinical environments remains constrained. Across the reviewed literature, diagnostic accuracy ranged between 50% and 70%, with ChatGPT-4 outperforming earlier iterations but still trailing human experts in reasoning, nuance, and contextual interpretation. These limitations reinforce the consensus that ChatGPT should be viewed not as an autonomous diagnostic agent but as a supportive tool within clinician-led workflows.

ChatGPT demonstrated moderate proficiency in standardized assessments in medical education, achieving pass rates between 61% and 67% on various national and speciality-specific exams. Learners and educators expressed optimism about its ability to generate personalized explanations, simulate test conditions, and support knowledge reinforcement. However, concerns about overreliance, deskilling, and the erosion of critical thinking skills persist.

These risks are particularly pronounced in high-stakes training environments like neurosurgery, where decision-making is non-linear and deeply context-sensitive.

Ethically, the review highlighted six dominant themes—accuracy, transparency, bias, accountability, overreliance, and governance—in over 70% of the included studies. The risk of AI hallucinations, opaque logic chains, and the lack of clear legal responsibility remain critical barriers to clinical trust and deployment. Moreover, only 17% of studies met low-risk criteria in bias assessment, revealing the need for more rigorous, reproducible, and ethically grounded research practices.

In summary, ChatGPT represents a transformative yet transitional tool. Its current utility lies in augmenting—not replacing—human expertise in surgical practice and pedagogy. The next research phase must focus on structured validation, responsible implementation, and ethical governance to fully realize its potential in neurosurgery.

Future Directions

To move beyond early-stage experimentation and conceptual optimism, the next phase of research on ChatGPT in neurosurgery must prioritize **rigorous, real-world validation**, **ethical safeguards**, and **broad generalizability**. The following strategic pathways are recommended:

1. **Multicenter, Real-World Clinical Validation**

   Future studies should assess ChatGPT's performance in authentic clinical settings across **multiple institutions and geographic regions**. These evaluations must use **real patient data**, integrate with **electronic health records (EHRs)**, and process **multimodal inputs** such as imaging, lab reports, and clinical notes. Cross-institutional collaboration will help reduce regional biases and improve generalizability.

2. **Longitudinal Impact Assessment**

   Rather than relying on isolated tests, longitudinal studies are needed to evaluate ChatGPT's sustained effects on clinical decision-making, diagnostic accuracy, workflow efficiency, and patient outcomes. In education, long-term exposure should be studied for its influence on knowledge retention, critical thinking, and independent reasoning among neurosurgical trainees.

3. **Development of Regulatory and Ethical Frameworks**

   Clear standards for liability, data privacy, transparency, and patient consent must accompany ChatGPT's deployment. Future work should support the creation of international and institutional **ethical governance models** that ensure accountability and prevent misuse in both clinical and educational environments.

4. **Adaptive Learning and Personalized Education**

   ChatGPT's potential as a dynamic learning assistant should be explored through **adaptive training tools** that personalize instruction based on specialty, trainee level, and learning style. Controlled studies with **feedback loops and performance analytics** are essential to evaluate pedagogical efficacy.

5. **Interdisciplinary Co-Design**

   To ensure safety and contextual relevance, AI development should be **co-designed** by neurosurgeons, AI developers, educators, ethicists, and patient representatives. This interdisciplinary approach will promote solutions that are both technologically sound and clinically grounded.

6. **Comparative Benchmarking with Other AI Models**

   With the rapid evolution of LLMs, comparative research is necessary to evaluate how ChatGPT performs relative to other platforms such as **Med-PaLM**, **Claude**, or domain-specific models. Key metrics should include diagnostic accuracy, transparency, user trust, and integration ease.

7. **Human Factors and Cognitive Load Research**

   ChatGPT's impact on clinicians' cognitive workload, decision-making efficiency, and trust must be studied using **human factors research**. Interface design should emphasize **traceability**, **usability**, and **explainability** to prevent automation bias and reduce decision fatigue.

Together, these directions form a roadmap for transitioning ChatGPT from an experimental tool to a validated, equitable, and ethically integrated system in neurosurgical practice and training. The emphasis must remain on amplifying human judgment, preserving educational integrity, and ensuring responsible innovation that serves diverse patient populations and learning communities.

REFERENCES

1. Abouammoh, N., Alhasan, K., Aljamaan, F., Raina, R., Malki, K. H., Altamimi, I., Muaygil, R., Wahabi, H., Jamal, A., Alhaboob, A., Assiri, R. A., Al-Tawfiq, J. A., Al-Eyadhy, A., Soliman, M., & Temsah, M.-H. (2025). Perceptions and earliest experiences of medical students and faculty with ChatGPT in medical education: Qualitative study. JMIR Medical Education, 11, e63400. https://doi.org/10.2196/63400
2. Abu Hammour, A., Abu Hammour, K., Alhamad, H., Nassar, R., El-Dahiyat, F., Sawaqed, M., Allan, A., Manaseer, Q., Abu Hammoura, M., Halboup, A., & Abu Farha, R. (2024). Exploring Jordanian medical students' perceptions and concerns about ChatGPT in medical education: A cross-sectional study. Journal of Pharmaceutical Policy and Practice, 17(1), 2429000. https://doi.org/10.1080/20523211.2024.2429000
3. Akindahunsi, T., Nwachukwu, C., & Amgbara, S. I. (2023). Neural decoding with artificial intelligence for personalized robotic neurosurgery. Open Access Research Journal of Science and Technology, 7(2), 40–48. https://doi.org/10.53022/oarjst.2023.7.2.0015
4. Ali, R., Tang, O. Y., Connolly, I. D., Zadnik Sullivan, P. L., Shin, J. H., Fridley, J. S., Asaad, W. F., Cielo, D., Oyelese, A. A., Doberstein, C. E., Gokaslan, Z. L., & Telfeian, A. E. (2023). Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery, 93(4), 470–477. https://doi.org/10.1227/neu.0000000000002632
5. Boaro, A., Mezzalira, E., Siddi, F., Bagattini, C., Gabrovsky, N., Marchesini, N., Broekman, M., & Sala, F. (2025). Knowledge, interest and perspectives on artificial intelligence in neurosurgery: A global survey. Brain and Spine, 5, 104156. https://doi.org/10.1016/j.bas.2024.104156
6. Bravo, J., Wali, A. R., Hirshman, B. R., Gopesh, T., Steinberg, J. A., Yan, B., Pannell, J. S., Norbash, A., Friend, J., Khalessi, A. A., & Santiago-Dieppa, D. (2022). Robotics and artificial intelligence in endovascular neurosurgery. Cureus, 14(3), e23662. https://doi.org/10.7759/cureus.23662
7. Dagi, T. F., Barker, F. G. II, & Glass, J. (2021). Machine learning and artificial intelligence in neurosurgery: Status, prospects, and challenges. Neurosurgery, 89(2), 133–142. https://doi.org/10.1093/neuros/nyab170
8. Dubinski, D., Won, S.-Y., Trnovec, S., Behmanesh, B., Baumgarten, P., Dinc, N., Konczalla, J., Chan, A., Bernstock, J. D., Freiman, T. M., & Gessler, F. (2024). Leveraging artificial intelligence in neurosurgery—unveiling ChatGPT for neurosurgical discharge summaries and operative reports. Acta Neurochirurgica. https://doi.org/10.1007/s00701-024-05908-3
9. El-Hajj, V. G., Gharios, M., Edström, E., & Elmi-Terander, A. (2022). Artificial intelligence in neurosurgery: A bibliometric analysis. World Neurosurgery. https://doi.org/10.1016/j.wneu.2022.12.087
10. Hong, D.-R., & Huang, C.-Y. (2024). The performance of AI in medical examinations: An exploration of ChatGPT in ultrasound medical education. Frontiers in Medicine, 11, 1472006. https://doi.org/10.3389/fmed.2024.1472006
11. Iqbal, J., Jahangir, K., Mashkoor, Y., Sultana, N., Mehmood, D., Ashraf, M., Iqbal, A., & Hafeez, M. H. (2022). The future of artificial intelligence in neurosurgery: A narrative review. Surgical Neurology International, 13, 536. https://doi.org/10.25259/SNI_877_2022

12. Khizar, A. (2023). An insight into artificial intelligence and its role in neurosurgery. Romanian Neurosurgery, 37(1), 124–127. https://doi.org/10.33962/roneuro-2023-021
13. Khizar, A. (2023). Artificial intelligence and neurosurgery: A revolution in the field. Pakistan Journal of Neurological Sciences, 18(4), Article 2. https://doi.org/10.56310/pjns.v18i04.244
14. Lawson McLean, A., & Gutiérrez Pineda, F. (2024). Application of transformer architectures in generative video modeling for neurosurgical education. International Journal of Computer Assisted Radiology and Surgery. Advance online publication. https://doi.org/10.1007/s11548-024-03266-0
15. Li, W., Gumera, A., Surya, S., Edwards, A., Basiri, F., & Eves, C. (2025). The role of artificial intelligence in diagnostic neurosurgery: A systematic review. Neurosurgical Review, 48, Article 393. https://doi.org/10.1007/s10143-025-03512-2
16. Mishra, R., & Deora, H. (2023). Artificial intelligence in neurosurgery: A review. Opinions in Medical Sciences, Technology and Health, 1(1), e23003. https://www.researchgate.net/publication/369235751
17. Mofatteh, M. (2021). Neurosurgery and artificial intelligence. AIMS Neuroscience, 8(4), 477–495. https://doi.org/10.3934/Neuroscience.2021025
18. Noh, S. H., Cho, P. G., Kim, K. N., Kim, S. H., & Shin, D. A. (2023). Artificial intelligence for neurosurgery: Current state and future directions. Journal of Korean Neurosurgical Society, 66(2), 113–120. https://doi.org/10.3340/jkns.2022.0130
19. Obande, J., Otobo, D. D., Alfin, J., & Shilong, D. (2025). Assessment of neurosurgical education in Nigeria: Looking at the preclinical and clinical educational foundation. Egyptian Journal of Neurosurgery, 40, Article 69. https://doi.org/10.1186/s41984-025-00396-8
20. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71
21. Shen, Q., Li, Z., Sun, Z., Liu, Y., Yang, S., Wang, X., Ji, L., & Zhang, W. (2024). ChatGPT and similar models in higher medical education of China: Refuse or embrace? Educational Research and Reviews, 6(7), 189–195. https://doi.org/10.32629/rerr.v6i7.2580
22. Sobhanian, P., Shafizad, M., Karami, S., Mozaffari, F., Arab, A., Razani, G., Shafiekhani, P., & Safari, S. (2022). Artificial intelligence applications in clinical neurosurgery. Precision Medicine and Clinical OMICS, 2(1), e133563. https://doi.org/10.5812/pmco-133563
23. Tangsrivimol, J. A., Schonfeld, E., Zhang, M., Veeravagu, A., Smith, T. R., Härtl, R., Lawton, M. T., El-Sherbini, A. H., Prevedello, D. M., Glicksberg, B. S., & Krittanawong, C. (2023). Artificial intelligence in neurosurgery: A state-of-the-art review from past to future. Diagnostics, 13(14), 2429. https://doi.org/10.3390/diagnostics13142429
24. Wójcik, S., Rulkiewicz, A., Pruszczyk, P., Lisik, W., Poboży, M., & Domienik-Karłowicz, J. (2024). Reshaping medical education: Performance of ChatGPT on a PES medical examination. Cardiology Journal, 31(3), 442–450. https://doi.org/10.5603/cj.97517
25. Zhou, J., Zhang, J., Wan, R., Cui, X., Liu, Q., Guo, H., Shi, X., Fu, B., Meng, J., Yue, B., Zhang, Y., & Zhang, Z. (2025). Integrating AI into clinical education: Evaluating general practice trainees' proficiency in distinguishing AI-generated hallucinations and impacting factors. BMC Medical Education, 25, 406. https://doi.org/10.1186/s12909-025-06916-2