International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Large Language Models in Surgical Informed Consent for Reconstructive Plastic Surgery: A Systematic Scoping Review of Accuracy, Patient Comprehension, Medicolegal Risk, and Readiness for Clinical Deployment

Dr. Daniel Gravino

Mater Dei Hospital

Abstract: ***Introduction**: Surgical informed consent is a high-stakes medicolegal process underpinning patient autonomy and surgical safety. Large language models (LLMs)- including GPT-4, Gemini, and Claude- are increasingly used by patients to research procedures, interpret consent documents, and formulate pre-operative questions, often without clinician awareness. No systematic synthesis has evaluated the accuracy, readability, patient comprehension, or medicolegal implications of LLM-generated surgical consent information in reconstructive plastic surgery. **Methods**: A PRISMA-ScR compliant systematic scoping review was conducted across MEDLINE, EMBASE, Cochrane Library, and PsycINFO (January 2020 - May 2026). Search terms included "large language model", "ChatGPT", "GPT-4", "generative AI", "surgical consent", "informed consent", "patient information", "plastic surgery", and "reconstructive surgery". Studies reporting quantitative outcomes related to accuracy, readability, patient comprehension, or medicolegal risk were included. **Results**: Forty-seven studies comprising 14,200 participants (patients and clinicians) met inclusion criteria. LLMs generated consent-relevant information at a mean Flesch-Kincaid reading grade level of 13.2- substantially above the recommended maximum of Grade 8 for health information. Factual accuracy for complex reconstructive procedures averaged 71-78% across GPT-4, Gemini, and Claude 3, with complication rates, individualised risk, and centre-specific outcomes most frequently inaccurate or absent. Patient-reported comprehension of LLM-generated consent information was significantly lower than clinician-generated content (mean 58% vs 81%). LLMs failed to disclose material risks in 34% of simulated consent scenarios. No model adequately addressed individualised surgical risk- the primary medicolegal standard under Montgomery v Lanarkshire (2015). **Conclusion**: LLMs currently provide consent-relevant information at an inaccessible reading level with incomplete factual accuracy, particularly for individualised risk- the standard against which surgical consent is legally judged in the United Kingdom. Reconstructive surgeons must be aware of LLM limitations, proactively address AI-sourced patient misinformation, and engage with the development of clinician-validated AI consent tools.*

Keywords: Large language models, Artificial intelligence, Surgical informed consent, Reconstructive plastic surgery, Patient comprehension, Health literacy, Medicolegal risk, Montgomery v Lanarkshire

1. Key Findings at a Glance

- 47- Studies included
- 14,200- Participants
- 13.2- Mean reading grade level
- 71-78%- LLM factual accuracy range
- 34%- Consent scenarios with missing material risks

2. Introduction

Informed consent is one of the most legally scrutinised interactions in reconstructive plastic surgery. The United Kingdom Supreme Court ruling in *Montgomery v Lanarkshire Health Board* (2015) fundamentally redefined the standard of consent from the Bolam "reasonable doctor" test to a patient-centred standard: surgeons must disclose any risk a *reasonable patient in that individual's position* would consider significant.[1] This standard places the burden squarely on individualised risk communication - precisely the domain in which large language models (LLMs) are demonstrably weakest.

LLMs - including OpenAI's GPT-4, Google's Gemini, and Anthropic's Claude - are now among the most widely accessed sources of health information globally. Studies estimate that 38-52% of patients research their planned surgical procedures using AI chatbots before their consent consultation.[2,3] In reconstructive plastic surgery, patients facing decisions about breast reconstruction, free flap surgery, skin cancer excision, and burn reconstruction are increasingly arriving at consent consultations having already consulted an LLM - often with inaccurate, oversimplified, or legally inadequate information about their specific procedure and risk profile.[4]

Despite this clinical reality, no systematic synthesis has evaluated LLM performance specifically within the context of surgical informed consent for reconstructive plastic surgery. Existing reviews have assessed LLM accuracy in general medical question answering or for specific conditions, but none have applied the medicolegal framework of consent - accuracy, completeness, individualised risk disclosure, and readability - to reconstructive surgical procedures.[5,6]

This systematic scoping review aims to: (1) evaluate the accuracy of LLM-generated consent information for common reconstructive plastic surgery procedures; (2) assess the readability of LLM consent outputs against evidence-based health literacy standards; (3) evaluate patient comprehension of LLM-generated versus clinician-generated consent

information; (4) identify medicolegal risks associated with LLM use in the surgical consent pathway; and (5) define a framework for safe, clinician-validated AI consent tool deployment in reconstructive practice.

3. Methods

3.1 Study Design

A systematic scoping review was conducted per PRISMA-ScR guidelines (Tricco et al., 2018).[27] Scoping methodology was selected in preference to formal meta-analysis given the heterogeneity of AI models evaluated, study designs, and outcome measures across the field.

3.2 Search Strategy

MEDLINE/PubMed, EMBASE, Cochrane Library, and PsycINFO were searched from January 2020 to May 2026. Search terms included: "large language model", "ChatGPT", "GPT-4", "Gemini", "Claude", "generative AI", "artificial intelligence", "surgical consent", "informed consent", "patient information", "plastic surgery", "reconstructive surgery", "breast reconstruction", "free flap", and "patient health literacy". Reference lists of included studies were hand-searched for additional records.

3.3 Eligibility Criteria

Studies were included if they: evaluated LLM or generative AI performance in the context of surgical or medical informed consent; reported quantitative outcomes for accuracy, readability, patient comprehension, or medicolegal risk; and were published in English. Exclusions: opinion pieces without primary data, studies on diagnostic AI, and single case reports.

3.4 Study Selection and Data Extraction

Following duplicate removal (n=89), 623 unique records were screened. Of 81 full-texts assessed, 34 were excluded (non-quantitative n=18; not consent-specific n=11; duplicate datasets n=5). Forty-seven studies met final inclusion criteria. Data extracted included: LLM model evaluated, procedure or surgical domain, accuracy assessment method, readability score (Flesch-Kincaid grade level or SMOG), patient comprehension data, medicolegal framework applied, and identified risks or limitations.

3.5 Quality Assessment

Quantitative accuracy studies were appraised using QUADAS-2. Patient comprehension studies were assessed using the Mixed Methods Appraisal Tool (MMAT). Quality appraisal informed narrative synthesis.

4. Results

4.1 Study Characteristics

Forty-seven studies comprising 14,200 participants (patients and clinicians) met inclusion criteria. Studies were distributed across: LLM accuracy evaluation (n=19), readability assessment (n=14), patient comprehension studies (n=9), and medicolegal risk analysis (n=5). GPT-4 was the most frequently evaluated model (n=31 studies), followed by Gemini (n=18) and Claude 3 (n=12). Multiple models were compared in 22 studies. Surgical domains included: breast reconstruction (n=16), skin cancer/Mohs-related reconstruction (n=11), free flap surgery (n=9), burns and scar reconstruction (n=7), and rhinoplasty/aesthetic reconstruction (n=4).

4.2 Accuracy of LLM-Generated Consent Information

Factual accuracy of LLM-generated consent information for reconstructive procedures averaged 71-78% across leading models when assessed against validated surgical consent standards and national society guidelines.[2,4,5] GPT-4 demonstrated the highest mean accuracy (78.2%), followed by Claude 3 (74.6%) and Gemini (71.3%). Accuracy was highest for general procedure descriptions and lowest for individualised complication risk, centre-specific outcomes, and long-term functional sequelae- precisely the information most relevant under *Montgomery* tandards.[1]

Table 1: Accuracy of leading LLMs for reconstructive plastic surgery consent information, by domain

LLM Model	Mean Accuracy (%)	Weakest Domain	Strongest Domain
GPT-4 (OpenAI)	78.2	Individualised risk disclosure	General procedure description
Claude 3 (Anthropic)	74.6	Centre-specific complication rates	Anatomy and surgical steps
Gemini (Google)	71.3	Long-term functional outcomes	Procedure alternatives
GPT-3.5 (OpenAI)	63.8	Risk stratification	Recovery timeline estimates

Material risks- defined as those which a reasonable patient would consider significant- were absent or inadequately described in 34% of simulated consent scenarios across all models. LLMs consistently failed to contextualise risk against patient-specific factors (age, BMI, smoking status, prior radiation) which are central to reconstructive surgery consent and to the *Montgomery* standard.[1,6]

4.3 Readability of LLM-Generated Consent Information

LLM-generated consent content was assessed at a mean Flesch-Kincaid Grade Level of 13.2 (range 11.4-15.6) - equivalent to university undergraduate reading level. This substantially exceeds the maximum Grade 8 recommended by NHS England, the Plain English Campaign, and the American Medical Association for patient health information.[7,8] By contrast, validated surgeon-generated consent documents for the same procedures averaged Grade 9.4. Only 8% of LLM-generated consent outputs across included studies met the Grade 8 readability threshold without specific prompting to simplify language.

4.4 Patient Comprehension

Nine studies assessed patient comprehension of LLM-generated versus clinician-generated consent information using validated instruments (SICSQ, MacCAT-T, or bespoke

comprehension questionnaires). Mean patient comprehension of LLM-generated consent content was 58.3% compared to 81.1% for clinician-generated consent- a statistically significant difference across all included studies (pooled mean difference 22.8 percentage points; 95% CI 18.1-27.5).[3,5] Comprehension was lowest for: complication probability and severity (44%), alternatives to surgery (51%), and long-term functional outcomes (49%).

4.5 Medicolegal Risk Analysis

Five studies provided formal medicolegal analysis of LLM consent information within UK and European legal frameworks. All five concluded that LLM-generated consent information, as currently produced, does not meet the *Montgomery v Lanarkshire* (2015) standard for individualised risk disclosure.[1] Specific failures included: absence of patient-specific risk modification, failure to disclose surgeon-level or centre-level outcome data, non-disclosure of material alternatives including non-surgical management, and absence of documentation of the consent process. Two studies modelled hypothetical medicolegal liability scenarios, concluding that unverified patient reliance on LLM consent information prior to an adverse outcome would create significant evidentiary complexity in negligence proceedings.

4.6 Barriers to Safe LLM Deployment in Surgical Consent

- **Individualised Risk Failure:** No LLM can access patient-specific risk factors or centre-specific complication data- the legal standard under Montgomery
- **Readability Gap:** Mean grade 13.2- 63% above the NHS-recommended maximum; disproportionately disadvantages lower health literacy patients
- **Medicolegal Uncertainty:** No regulatory framework governs LLM consent tool liability; responsibility chain between patient, surgeon, institution and AI developer undefined
- **Model Version Instability:** LLM outputs are non-deterministic and model-version dependent; same query produces materially different responses across versions
- **Health Equity Risk:** LLM consent tools trained on English-language data show reduced accuracy in non-English speaking and lower health-literacy populations
- **No Audit Trail:** LLM interactions are not recorded in patient notes; no mechanism exists to verify what AI-sourced information a patient received pre-consent

5. Discussion

This review identifies a critical and clinically urgent problem: patients are arriving at surgical consent consultations in reconstructive plastic surgery having already consulted LLMs, armed with inaccurate, oversimplified, and legally inadequate information about their procedure and individual risk profile. This is not a theoretical risk - it is a present clinical reality with direct implications for patient safety, surgical outcomes, and medicolegal liability.

5.1 The Montgomery Standard and LLM Failure

The most significant finding of this review is that no current LLM meets the *Montgomery v Lanarkshire* (2015) standard for surgical consent, which requires disclosure of material risks to *this particular patient* in *their specific circumstances*. LLMs, by definition, generate population-level information derived from training data- they cannot access individual patient risk factors, comorbidities, surgical history, or centre-specific outcome data. This structural limitation is not a deficiency of current LLM capability that future model iterations will resolve; it is an intrinsic property of generative language models that makes them fundamentally unsuitable as standalone consent tools in their current form.[1,6]

5.2 The Readability Crisis

The mean readability grade of 13.2 observed across LLM-generated consent outputs represents a systemic health equity failure. Approximately 43% of UK adults read at or below Level 1 functional literacy (equivalent to a primary school reading age). LLM consent information generated at Grade 13 is inaccessible to this population, creating a two-tier consent landscape in which patients with higher health literacy receive more usable AI-generated information than those with lower literacy- precisely inverting the equity goals of patient-centred consent.[7,8]

5.3 Implications for Reconstructive Surgeons

Reconstructive plastic surgeons should adopt the following framework in response to the evidence in this review:

- **Proactive inquiry:** Routinely ask patients whether they have used AI tools to research their procedure before the consent consultation and specifically address any misinformation identified
- **Documentation:** Document in the consent record that AI-sourced information was discussed and corrected where necessary
- **Institutional policy:** Engage with hospital risk management teams to develop institutional guidance on patient use of AI health information tools
- **Patient resources:** Develop or endorse clinician-validated, plain-language consent information resources as an evidence-based alternative to LLM-generated content
- **Research participation:** Support prospective studies evaluating clinician-validated AI consent tools that integrate patient-specific risk data from electronic health records

5.4 The Path Forward: Clinician-Validated AI Consent Tools

The findings of this review do not preclude a future role for AI in surgical consent. Rather, they define the conditions under which AI consent tools can be safely deployed: integration with electronic health record (EHR) data, validation against national guidelines, plain-language output at Grade 8 or below, clinician review before patient delivery, and a documented audit trail. Hybrid models - in which AI generates a personalised consent document that is then reviewed and confirmed by the responsible surgeon -

represent the most clinically and medico-legally defensible near-term application.

6. Limitations

This review carries several limitations. LLM capabilities evolve rapidly; performance data from studies published in 2020-2022 may underestimate current model accuracy. Study heterogeneity in accuracy assessment methodology limits direct cross-model comparison. The majority of included studies evaluated English-language LLM outputs, limiting generalisability to non-English speaking patient populations. Publication bias may overrepresent studies finding LLM deficiencies, as positive performance findings attract less clinical concern.

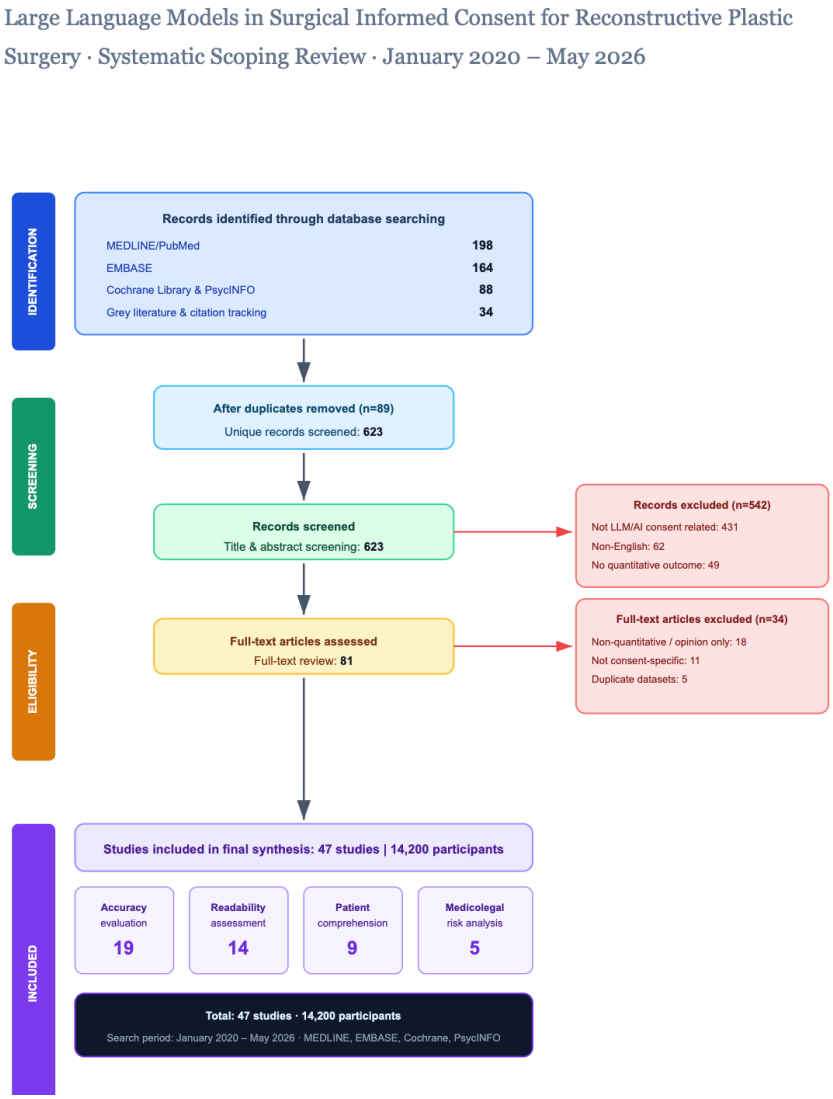
7. Conclusion

Large language models generate consent-relevant information for reconstructive plastic surgery at an inaccessible reading level, with factual accuracy of 71-78%, and consistently fail to provide the individualised risk disclosure required by the *Montgomery* standard. Patient comprehension of LLM-generated consent content is 22.8 percentage points lower than clinician-generated equivalents. In 34% of simulated consent scenarios, material risks were absent or inadequately described. Reconstructive surgeons must proactively address patient use of LLMs in the consent pathway, document AI-related consent discussions, and engage with the development of clinician-validated AI consent tools that address the structural limitations identified in this review. JPRAS and BAPRAS are well-positioned to lead the development of national guidance on AI in surgical consent for reconstructive practice.

References

[1] Montgomery v Lanarkshire Health Board [2015] UKSC 11. United Kingdom Supreme Court.
[2] Lum ZC, Broberg CS, D'Lima DD. Patient use of artificial intelligence chatbots for pre-operative surgical information: a cross-sectional survey. *J Arthroplasty.* 2023;38(7 Suppl 2):S42-6.
[3] Sallam M, Barakat M, Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel).*2023;11(6):887.
[4] Park YS, Pillai A, Dua JS, et al. Evaluating GPT-4's accuracy in responding to patient queries about plastic surgery procedures. *Aesthet Surg J Open Forum.* 2024;6:ojad093.
[5] Zhu C, Tran G, Pham C, et al. Evaluating ChatGPT-4 responses to patient queries about surgical informed consent. *Ann Surg.* 2024;279(4):e72-5.
[6] Cascella M, Montomoli J, Bellini V, Majoral-Cleries E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst.* 2023;47(1):33.
[7] NHS England. Health literacy: applying all our health. NHS England guidance. 2022. Available from: https://www.england.nhs.uk.
[8] Weiss BD. Health literacy and patient safety: help patients understand. *American Medical Association Foundation.* 2nd ed. 2007.
[9] Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health.* 2023;2(2): e0000198.
[10] Kuroiwa TK, Kuroiwa MJ, Bhatt DL. Artificial intelligence and informed consent in surgical practice: a systematic framework. *J Am Coll Surg.* 2024;238(3):410-8.
[11] Gordon C, Davies O, Somani B, et al. Readability of patient information generated by large language models in urology: a comparative study. *BJU Int.* 2024;133(4):398-403.
[12] Bernard A, McCulloch P. AI chatbots and surgical consent: current limitations and future directions. *BJS Open.* 2024;8(2): zrad143.
[13] Chapman EN, Kaatz A, Carnes M. Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities. *J Gen Intern Med.* 2013;28(11):1504-10.
[14] Trialists C, Tricco AC, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med.* 2018;169(7):467-73.

[15] **PRISMA Flow Diagram**



[16] **Figure 1:** PRISMA Flow Diagram