

Review Article

Artificial intelligence in plastic and reconstructive surgery: clinical applications, translational readiness and barriers to real-world implementation

Sergio Miguel Suárez Fuentes^{1*}, Mauro Axel Cárdenas Rosales², Andrea Nallely Garza Cisneros³, Martin Sergio Rodríguez Garza³, Luis Fernando Hernández González⁴, Valeria Gómez Márquez¹, Nora Lis Flores Olmos¹

¹Department of General Surgery, Hospital Regional Dr. Valentín Gómez Farías, ISSSTE, Zapopan, Jalisco, Mexico

²Department of General Surgery, Hospital General Presidente Lázaro Cárdenas del Río, ISSSTE, Chihuahua, Chihuahua, Mexico

³Department of General Surgery, Hospital Regional ISSSTE Monterrey, Monterrey, Nuevo León, Mexico

⁴Universidad de Guadalajara, Guadalajara, Jalisco, Mexico

Received: 07 September 2026

Revised: 12 October 2026

Accepted: 19 October 2026

*Correspondence:

Dr. Sergio Miguel Suárez Fuentes,
E-mail: sergiomsuarez@gmail.com

Copyright: © the author(s), publisher and licensee Medip Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

Artificial intelligence (AI) is expanding rapidly in plastic and reconstructive surgery, but the central question has shifted from whether an algorithm can perform a task to whether it can generalize, improve clinical decisions, and remain safe after deployment. This critical review examines contemporary AI applications in burn assessment, facial and craniofacial surgery, microsurgical free-flap monitoring, breast reconstruction, and patient communication, with emphasis on translational readiness rather than cataloguing technologies. A structured PubMed/MEDLINE search through September 6, 2026 was supplemented by reference chaining from recent systematic reviews and reporting guidance. Image-based systems have reached high performance for selected bounded tasks, including automated free-flap monitoring (area under the curve 0.960), burn-depth assessment (95% accuracy in a prospective clinical cohort), and some externally validated breast-reconstruction outcome models (area under the curve up to 0.97). In contrast, prediction of multifactorial complications remains less discriminative, and general-purpose large language models have shown clinically important limitations in factual accuracy and postoperative triage. The principal barrier is therefore not algorithm development but clinical translation: external validation, prospective impact studies, subgroup performance, human-AI workflow design, data governance, and post-deployment monitoring remain inconsistently addressed. Plastic surgery is well positioned to benefit from AI because it is image-rich and outcome-sensitive, but implementation should prioritize narrowly defined tasks with measurable clinical value and preserve surgeon oversight, patient-reported outcomes, and transparent lifecycle governance.

Keywords: Artificial intelligence, Plastic surgery, Reconstructive surgery, Machine learning, Clinical implementation

INTRODUCTION

Artificial intelligence (AI) has moved from experimental pattern recognition to a broad family of tools capable of classification, segmentation, prediction, natural-language

generation, and multimodal data integration. In surgical practice, machine learning and deep learning can combine clinical variables, imaging, and perioperative data to support decisions that are difficult to standardize by human assessment alone. Previous work across general surgery

has identified risk prediction, operative assistance, perioperative decision support, and outcome optimization as major domains of AI application, while also emphasizing the persistent gap between technical performance and real-world clinical implementation.¹ Plastic and reconstructive surgery is an especially favorable environment for these methods because clinical photography, three-dimensional imaging, perfusion assessment, anthropometry and patient-reported outcomes are integral to the specialty.

The literature has expanded quickly. Broad reviews have catalogued AI applications across plastic surgery, facial aesthetics, aesthetic surgery, and microsurgery.²⁻⁵ A 2026 systematic review focused on facial aesthetic surgery again confirmed rapid growth but also emphasized that much of the evidence remains concentrated in development, feasibility, and validation rather than interventional clinical evaluation.⁶ Consequently, another catalogue of AI applications would add limited value. The more clinically relevant question is which applications have crossed the translational gap from technically impressive performance to evidence that is generalizable, clinically actionable, and deployable.

This distinction matters because internal test-set performance can overstate real-world utility. Dataset shift, differences in skin pigmentation, photographic devices, lighting, surgical technique, outcome prevalence, and local practice patterns can materially alter model behavior. Contemporary reporting and implementation frameworks such as TRIPOD+AI, CLAIM, and FUTURE-AI therefore emphasize external validation, transparent reporting, fairness, robustness, human factors, and lifecycle monitoring.⁷⁻⁹ This review examines the strongest contemporary applications of AI in plastic and reconstructive surgery through that translational lens and identifies the evidence required before routine clinical implementation.

LITERATURE SEARCH AND REVIEW APPROACH

This article was designed as a critical, structured narrative review rather than a systematic review or meta-analysis. PubMed/MEDLINE was searched for English-language literature available through September 6, 2026 using combinations of the terms artificial intelligence, machine learning, deep learning, large language model, ChatGPT, plastic surgery, reconstructive surgery, aesthetic surgery, breast reconstruction, microsurgery, free flap, burn, craniofacial, and orthognathic surgery. Reference lists of recent systematic reviews were used for backward citation searching.²⁻⁶ Priority was given to peer-reviewed original studies from 2021 onward that reported quantitative clinical performance, external validation, prospective evaluation, or implementation, together with major reporting and deployment guidance.⁷⁻⁹ Earlier or nonclinical publications were retained only when necessary to establish context. Because the objective was

interpretive synthesis of heterogeneous technologies and clinical tasks, no pooled effect estimate was calculated and the studies summarized in the tables should be considered representative rather than an exhaustive denominator of all AI publications in the specialty.

WHY TRANSLATIONAL READINESS IS THE KEY QUESTION

A useful distinction in medical AI is the difference between technical validity and clinical utility. Technical validity asks whether a model can reproduce a label or predict an outcome in a defined dataset. Clinical utility asks whether the model remains calibrated in a new population, changes clinician behavior appropriately, improves a patient-relevant outcome, reduces workload or cost, and does so without introducing inequity or unsafe automation. A model can achieve an area under the receiver operating characteristic curve (AUC) above 0.90 and still fail clinically if the test population resembles the training cohort too closely, if the output is not actionable, or if false reassurance carries substantial harm.⁷⁻⁹

This problem is visible in plastic surgery because AI tasks range from tightly bounded image classification to highly contextual decisions. Detecting a color or temperature change in an exposed free flap is a substantially narrower problem than predicting flap loss from preoperative variables, and generating educational text is fundamentally different from deciding whether a postoperative patient requires emergency reassessment. Translational readiness should therefore be judged against the clinical task, intended user, consequence of error, and degree of human oversight rather than a single performance metric. Figure 1 presents a practical pathway from model development to sustained clinical deployment.

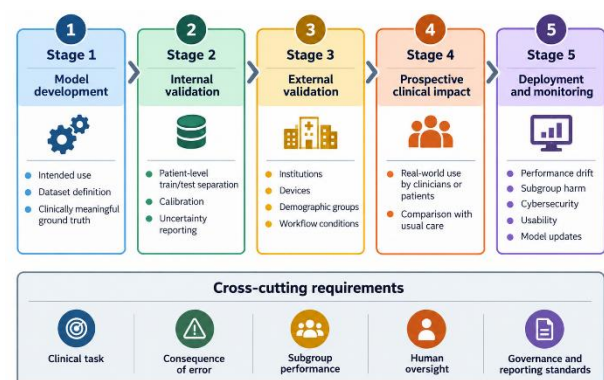


Figure 1: Translational pathway for artificial intelligence in plastic and reconstructive surgery. The pathway emphasizes that model development and internal performance are early stages; external validation, prospective clinical impact, and post-deployment monitoring are required before routine use. At every stage, the intended clinical task, consequences of error, subgroup performance, human oversight, and governance should be explicit.

AI IN BURN ASSESSMENT

Burn care is a natural target for computer vision because management depends on visually derived variables such as wound extent and depth. Chang et al demonstrated deep-learning segmentation of burn wounds from clinical images, supporting automated measurement of injured surface area.¹⁰ A subsequent study using 4,991 burn-wound images and 1,050 palm images reported precision of 0.908 and sensitivity of 0.901 for total-wound segmentation with DeepLabV3+, while palm segmentation approached 0.99.¹¹ These studies illustrate strong technical performance for a clearly defined visual task.

The translational step is more important. El Masry et al developed AMBUSH-AI, integrating B-mode ultrasound, tissue Doppler elastography, and AI interpretation to classify burn depth. After preclinical development, prospective clinical evaluation in 30 patients achieved 95% accuracy for identifying third-degree burns.¹² This moves beyond retrospective image classification toward a decision with direct implications for excision and grafting. Nevertheless, broader deployment requires multicenter evaluation across skin tones, injury mechanisms, operators, and acquisition hardware. For burn AI, the most credible path to implementation is therefore not replacement of expert assessment, but standardized decision support in settings where burn expertise is limited or interobserver variability is high.

FACIAL, AESTHETIC AND CRANIOFACIAL APPLICATIONS

Facial plastic surgery has generated a large AI literature because standardized photography, cephalometry, and three-dimensional imaging allow quantitative analysis of shape and symmetry. Zhang et al used four neural networks in 50 facelift patients and found an AI-estimated apparent-age reduction of 4.3 years compared with 6.7 years by patient self-appraisal; AI-derived age change correlated with FACE-Q satisfaction.¹³ Lo et al used three-dimensional contour features in 158 orthognathic patients and reported improvement in mean symmetry score from 2.74 to 3.52 after surgery.¹⁴

AI has also been evaluated as a decision-support tool for orthognathic treatment. Shin et al. analyzed cephalograms from 840 individuals and reported test accuracy of 0.954, sensitivity of 0.844, and specificity of 0.993 for predicting the need for orthognathic surgery.¹⁵ In a separate cohort of 920 radiographs, de Oliveira et al compared ten machine-learning models; the combined approach achieved AUC 0.791 overall, 0.824 in Class II patients, and 0.890 in Class III patients.¹⁶ Ho et al later combined three-dimensional contour maps with hyperdimensional computing, training on 5,150 images and clinically evaluating 200 patients; the system achieved 84.1% accuracy and quantified a 53% postoperative improvement in symmetry score.¹⁷

The most important limitation is constructing validity. In facial feminization surgery, automated female classification increased from 48.1% preoperatively to 74.1% postoperatively, but algorithmic classification was not equivalent to patient satisfaction.¹⁸ Similarly, in 48 Chinese women after facelift surgery, AI estimated apparent-age reduction differently from both blinded observers and the patients themselves.¹⁹ These studies demonstrate that reproducible computational metrics can be clinically informative without being valid substitutes for patient-defined success. Aesthetic AI should therefore complement, not displace, patient-reported outcome measures and shared decision-making. It also requires explicit testing across ethnic and phenotypic groups because training distributions may encode culturally narrow standards of age, symmetry, femininity, or attractiveness.

MICROSURGERY AND FREE-FLAP MONITORING

Microsurgery provides one of the clearest examples of how task definition determines AI performance. Asaad et al. developed machine-learning models in 4,000 head-and-neck free-flap reconstructions. Any complication occurred in 33.7%, major recipient-site complications in 26.5%, and total flap loss in 1.7%; the best AUCs were 0.61, 0.68, and 0.66, respectively.²⁰ These modest values reflect the difficulty of predicting multifactorial outcomes from preoperative and perioperative variables.

By contrast, models focused on direct perfusion assessment have performed substantially better. Huang et al analyzed 805 postoperative flap observations and reported 98.4% accuracy with a random-forest model distinguishing normal perfusion, arterial insufficiency, and venous insufficiency; temperature and color differences were dominant predictors.²¹ Kim et al developed an automated smartphone-based monitoring system using photographs from 305 patients. The segmentation model achieved a Dice coefficient of 0.970; the perfusion classifier, trained on 11,112 images, achieved AUC 0.960, with sensitivity of 97.5% for venous insufficiency and 92.8% for arterial insufficiency. Crucially, the integrated system was subsequently used in ten patients across 143 automated monitoring sessions.²²

This sequence - objective signal, high internal performance, integration into a clinical workflow, and initial prospective use - is closer to genuine translational readiness than many purely retrospective prediction studies. The remaining questions are external: performance across different skin pigmentation, flap types, lighting conditions, smartphone cameras, buried flaps, and institutions. Automated monitoring is therefore best conceptualized as a sentinel system that can escalate concern to clinicians rather than as an autonomous replacement for bedside examination.

Table 1: Representative clinical evidence illustrating different levels of AI translation in plastic and reconstructive surgery.

Domain	Study	Population / data	AI task	Reported performance	Translation status
Burns	Chang et al ¹¹	4,991 wound images + 1,050 palm images	Automated segmentation / %TBSA support	Precision 0.908; sensitivity 0.901 for total wound	Internal image validation
Burns	El Masry et al ¹²	30 patients (prospective clinical phase)	Burn-depth classification	95% accuracy for third-degree burns	Prospective clinical evaluation
Facial / orthognathic	Shin et al ¹⁵	840 patients	Need for orthognathic surgery	Accuracy 0.954; sensitivity 0.844; specificity 0.993	Internal validation
Facial / orthognathic	de Oliveira et al ¹⁶	920 radiographs	Surgery vs orthodontic treatment	AUC 0.791 overall; 0.890 in Class III	Train/validation/test split
Microsurgery	Asaad et al ²⁰	4,000 free-flap reconstructions	Postoperative complication prediction	AUC 0.61-0.68 depending on outcome	Retrospective internal validation
Microsurgery	Huang et al ²¹	805 flap observations	Perfusion-state classification	98.4% accuracy	Retrospective internal validation
Microsurgery	Kim et al ²²	305 patients; 11,112 images; 10-patient deployment	Automated flap monitoring	AUC 0.960; venous sensitivity 97.5%	Integrated clinical deployment pilot
Breast reconstruction	Ozmen et al ²⁵	13,312 DIEP procedures	30-day readmission	AUC 0.8921; accuracy 88%; recall 79%	Large database internal test set
Breast reconstruction	Gabay et al ²⁶	27,718 patients	Readmission, reoperation, prolonged stay	AUC 0.614-0.861	Large database internal validation
Breast reconstruction	Chen et al ²⁷	2,687 developments + 2,089 external validations	1-year BREAST-Q improvement	AUC up to 0.97	External multicenter validation
Generative AI	Hung et al ²⁸	Patient education materials	Breast-reconstruction education	50% of ChatGPT content judged accurate	Expert comparison; no clinical deployment
Generative AI	Abi-Rafeh et al ²⁹	22 simulated complications	Postoperative triage	44% most-likely diagnosis; 55% disposition	Simulation only

BREAST RECONSTRUCTION AND INDIVIDUALIZED OUTCOME PREDICTION

Large reconstructive datasets allow AI to identify risk phenotypes and predict outcomes that are difficult to estimate at the individual level. Using 14,274 patients from the National Surgical Quality Improvement Program

(NSQIP), Kim et al applied unsupervised learning and identified seven clinically distinct subgroups after autologous breast reconstruction, with complication profiles ranging from relatively low-risk to clusters enriched for reoperation, readmission, and wound dehiscence.²³ In implant-based reconstruction, Chen et al analyzed 209 patients (406 breasts); a neural-network

model using clinically significant variables achieved an AUC of 0.74 for secondary procedures, with sensitivity of 84.2%.²⁴

More recent national-database models have improved discrimination. Ozmen et al evaluated 13,312 DIEP flap procedures and reported a 5.42% readmission rate; a stacked model achieved 88% accuracy, 79% recall, and AUC 0.8921 for 30-day readmission.²⁵ Gabay et al analyzed 27,718 breast reconstructions across autologous, implant, and tissue-expander cohorts; AUCs ranged from 0.614 to 0.861, with the strongest performance for prolonged length of stay after implant reconstruction and readmission after delayed autologous reconstruction.²⁶

The strongest evidence of generalizability comes from external validation. Chen et al. developed models using 2,687 patients and externally validated them in 2,089 patients from the Mastectomy Reconstruction Outcomes Consortium. For one-year BREAST-Q improvement, reported AUCs reached 0.97 for abdominal physical well-being and remained clinically informative across breast satisfaction, sexual well-being, chest physical well-being, and psychosocial well-being.²⁷

This is a meaningful advance because the target outcome is patient-reported and the model was tested outside the development cohort. The next translational step is prospective decision-impact evaluation: whether providing these predictions to surgeons and patients improves counseling, procedure selection, expectation alignment, or postoperative resource use.

GENERATIVE AI AND PATIENT COMMUNICATION

Large language models (LLMs) require a different safety framework because fluent language can obscure factual error. Hung et al compared breast-reconstruction education materials produced by experts and ChatGPT. Expert materials were more readable (approximately grade 7.5 versus grade 10.5), and only 50% of ChatGPT-generated content was judged accurate.²⁸ Abi-Rafeh et al tested Bard in 22 simulated postoperative body-contouring complications. After history-taking, the model identified the most likely diagnosis correctly in 44% of scenarios, suggested the indicated disposition in 55%, and recognized red-flag features with 48% accuracy.²⁹ These performance levels are inadequate for autonomous postoperative triage.

Domain specialization can improve language-model behavior without proving clinical safety. PlasticSurgeryGPT was fine-tuned on 25,389 plastic-surgery abstracts and outperformed its baseline GPT-2 model on BLEU, METEOR, and ROUGE-1 linguistic metrics.³⁰ However, these metrics measure text similarity rather than diagnostic safety, guideline adherence, or patient outcomes. In 2026, Perez Rivera et al compared ChatGPT-5 with a medical evidence-oriented platform across 20 aesthetic procedures. OpenEvidence produced higher DISCERN scores and greater citation accuracy, while readability and actionability were similar.³¹ This finding supports retrieval-grounded or evidence-linked systems for patient education, but it does not justify unsupervised clinical decision-making.

For current practice, LLMs are most defensible for low-risk functions such as drafting educational material, summarizing information for clinician review, or administrative assistance. High-risk outputs - postoperative triage, complication diagnosis, procedural eligibility, or recommendations that could delay urgent evaluation - require explicit human verification and conservative escalation rules.

A PRACTICAL TRANSLATIONAL-READINESS FRAMEWORK

The evidence can be organized into five stages rather than a binary distinction between experimental and clinical AI. Stage 1 is model development with a clearly defined intended use and clinically meaningful ground truth. Stage 2 is internal validation with patient-level separation between training and test data, calibration, and appropriate uncertainty reporting. Stage 3 is external validation across institutions, devices, demographic groups, and workflow conditions. Stage 4 is prospective clinical impact evaluation in which the model is actually used by clinicians or patients and compared with usual care. Stage 5 is deployment with ongoing monitoring for performance drift, subgroup harm, cybersecurity, usability, and model updates.⁷⁻⁹

Most plastic-surgery AI remains in stages 1-2. The strongest examples approaching stages 3-4 include prospective burn-depth assessment, automated free-flap monitoring with initial clinical deployment, and externally validated breast-reconstruction outcome prediction.^{12,22,27} Table 2 translates these stages into evidence requirements that can be applied by surgeons, reviewers, and institutions when evaluating a new tool.

Table 2: Practical translational-readiness framework for clinical AI in plastic surgery.

Stage	Minimum evidence	Principal unresolved question	Illustrative status in current literature
1. Development	A clinically bounded intended use, representative ground truth, patient-level data separation	Technical discrimination alone	Burn segmentation; domain-specific LLM proof of concept

Continued.

Stage	Minimum evidence	Principal unresolved question	Illustrative status in current literature
2. Internal validation	Independent test set from development environment, calibration, uncertainty, error analysis	Generalization outside the source institution	Most retrospective prediction and imaging studies
3. External validation	Independent institutions/populations/devices; subgroup and calibration analysis	Clinical impact when used in practice	Externally validated breast-reconstruction outcome prediction
4. Prospective impact	Human-plus-AI workflow compared with usual care; patient and workflow outcomes	Durability and model drift	Prospective burn assessment; initial automated flap monitoring
5. Deployment and monitoring	Governance, version control, cybersecurity, drift monitoring, update pathway, audit trail	Ongoing safety and equity	Target state for routine clinical AI

GENERALIZABILITY, FAIRNESS AND HUMAN FACTORS

Plastic surgery is unusually sensitive to dataset representativeness. Skin pigmentation affects visual cues in burns and flap monitoring; facial morphology and aesthetic preferences vary across ancestry, age, sex, gender identity, and culture; and reconstructive outcomes depend on institutional technique and thresholds for reoperation. External validation should therefore be planned around clinically plausible sources of shift rather than performed as a token split of the same database.⁷⁻⁹

Fairness analysis should report clinically relevant subgroup performance, but equal performance metrics alone are insufficient. The consequences of error matter. Missing venous congestion in a free flap, underestimating a deep burn, or falsely reassuring a patient with a postoperative complication carries a different risk profile from a small error in apparent-age estimation. Human factors are equally important: alerts must be interpretable, actionable, and integrated into workflow without creating alarm fatigue or automation bias. The relevant unit of performance is often the human-AI team, not the algorithm alone.⁹

REGULATION, GOVERNANCE AND MODEL LIFECYCLE

Clinical deployment also requires governance beyond publication. Good Machine Learning Practice principles emphasize representative datasets, independent test sets, clinically relevant performance, transparency, human-AI team performance, and monitoring of deployed models.³² For software that functions as a medical device, iterative updates create an additional challenge: performance may change after deployment as data distributions, hardware, and model versions evolve. Institutions should therefore document the intended use, model version, validation

population, update process, fallback pathway, and responsibility for acting on outputs.

Data governance is particularly important in plastic surgery because facial and body photographs may be identifiable biometric data. Uploading clinical images to external AI services without an appropriate legal basis, security controls, and institutional governance can create confidentiality risks independent of model accuracy. Local deployment decisions should therefore involve surgery, information security, privacy, clinical engineering, and ethics rather than being treated solely as a software procurement decision.

IMPLICATIONS FOR LATIN AMERICAN PRACTICE

AI could be valuable in Latin American reconstructive practice precisely where specialist resources are unevenly distributed. Smartphone-based flap monitoring, standardized burn assessment, and validated risk tools could support regional hospitals and postoperative follow-up if they are designed for constrained environments.

However, importing models trained predominantly in North American or East Asian cohorts without local validation risks systematic error. Multicenter Latin American datasets should capture skin-tone diversity, imaging hardware, socioeconomic and access variables, local operative techniques, and patient-reported outcomes in Spanish and other regional languages.

The most realistic implementation strategy is selective rather than comprehensive: begin with a bounded, high-value task; validate it locally; define an escalation pathway; measure workload and patient outcomes prospectively; and monitor performance after deployment. This approach is more likely to produce durable clinical benefit than adopting broad generative systems because they are widely available.

Table 3: Priority implementation questions by clinical domain.

Domain	Near-term opportunity	Key implementation risk	Most useful next study
Burn assessment	Standardize wound extent/depth; support nonspecialist triage	Skin tone, image quality, operator/device variability	Multicenter prospective diagnostic-impact study
Facial / craniofacial	Quantitative symmetry; planning; outcome measurement	Construct validity, phenotypic/cultural bias, surrogate outcomes	External validation linked to patient-reported outcomes
Free-flap monitoring	Continuous sentinel monitoring and early escalation	Lighting, skin pigmentation, camera models, buried flaps, alarm fatigue	Multicenter prospective human-plus-AI salvage/time-to-detection study
Breast reconstruction	Individualized risk and patient-reported outcome prediction	Institutional technique, calibration, treatment-selection effects	Prospective decision-impact trial with calibration monitoring
Generative AI	Education, documentation support, evidence synthesis	Hallucination, citation errors, outdated versions, unsafe triage	Retrieval-grounded systems with clinician review and safety escalation

DISCUSSION

The contemporary evidence supports a nuanced conclusion. AI is already capable of performing several plastic-surgery tasks at a level that is potentially useful to clinicians, but the maturity of the evidence varies sharply by application. Computer vision and structured prediction perform best when the target is well defined: burn segmentation, perfusion classification, facial measurement, and selected breast-reconstruction outcomes. In contrast, broad complication prediction and general-purpose conversational AI remain more vulnerable to contextual complexity, low event rates, hallucination, and unsafe overconfidence. This pattern helps reconcile two apparently conflicting observations in the literature: plastic surgery has many AI publications, yet relatively few tools have become routine clinical infrastructure.²⁻⁶ The bottleneck is no longer proof that AI can recognize patterns. It is evidence that a specific model, used by a specific person in a defined workflow, improves a decision or outcome in a population different from the one used to build it. External validation and prospective impact studies should therefore carry greater weight than marginal improvements in internal AUC.

The strongest near-term opportunity is augmentation rather than autonomy. A free-flap system can continuously screen photographs and alert the surgical team; a burn system can provide a standardized second opinion; a breast-reconstruction model can make risk or patient-reported outcome probabilities more explicit during counseling. In each example, the surgeon retains responsibility for contextual interpretation. This design is consistent with contemporary trustworthy-AI guidance and reduces the risk that automation silently substitutes for clinical judgment.⁷⁻⁹ Future studies should report calibration, decision thresholds, false-negative consequences, subgroup performance, missing-data handling, and model availability in addition to

discrimination. Multicenter validation should precede broad clinical claims. Prospective trials should compare human-alone versus human-plus-AI workflows and include patient-relevant outcomes, time burden, cost, and unintended harms. For generative systems, evidence retrieval, citation traceability, version control, and explicit uncertainty or escalation behavior should be treated as safety requirements rather than optional features.³¹

Finally, the specialty should resist using AI-derived aesthetic metrics as surrogate endpoints without validation against patient priorities. The discrepancy between algorithmic measures and patient satisfaction in facial surgery demonstrates why technical objectivity does not automatically equal clinical validity.^{18,19} Patient-reported outcomes and shared decision-making remain central, particularly where treatment goals are preference-sensitive.

Limitations

This review has limitations. It is a structured narrative review rather than a systematic review, and the search was centered on PubMed/MEDLINE with reference chaining; therefore, it was not designed to identify every AI publication in plastic surgery. The field is highly heterogeneous, spanning images, clinical variables, three-dimensional data, text generation, and different outcome definitions, which precludes meaningful pooled performance estimates.

Commercial LLMs and some device software change rapidly, so published evaluations are version-specific and may not represent later releases. Finally, the translational-readiness framework proposed here is a practical interpretive model informed by contemporary reporting and deployment guidance rather than a validated scoring instrument.

CONCLUSION

Artificial intelligence has credible clinical potential in plastic and reconstructive surgery, particularly for image-based assessment, free-flap monitoring, burn evaluation, craniofacial measurement, and individualized outcome prediction. The field should now shift its emphasis from demonstrating algorithmic performance to demonstrating clinical translation. External multicenter validation, prospective human-AI impact studies, subgroup and fairness analysis, secure data governance, and post-deployment monitoring are the decisive next steps. General-purpose LLMs remain unsuitable for autonomous postoperative triage or other high-risk decisions without clinician verification. The most defensible implementation model is task-specific augmentation: AI should make relevant signals more measurable, timely, and individualized while preserving surgeon oversight and patient-defined outcomes.

Funding: No funding sources

Conflict of interest: None declared

Ethical approval: Not required

REFERENCES

- Suárez-Fuentes SM, Flores-Chávez LA, Martínez-López EM, Castrejón-Cardona CD, Menéndez-Gutiérrez LL. The impact of artificial intelligence in general surgery: enhancing precision, efficiency, and outcomes. *Int J Res Med Sci.* 2025;13(1):293-7.
- Kiwan O, Al-Kalbani M, Rafie A, Hijazi Y. Artificial intelligence in plastic surgery, where do we stand? *JPRAS Open.* 2024;42:234-43.
- Stephanian B, Karki S, Debnath K, Saltychev M, Rossi-Meyer M, Kandathil CK, et al. Role of artificial intelligence and machine learning in facial aesthetic surgery: a systematic review. *Facial Plast Surg Aesthet Med.* 2024;26(6):679-705.
- Nogueira R, Eguchi M, Kasmirski J, Veronez de Lima B, Dimatos DC, Lima DL, et al. Machine learning, deep learning, artificial intelligence and aesthetic plastic surgery: a qualitative systematic review. *Aesthetic Plast Surg.* 2025;49(1):389-399.
- Savage SA, Seth I, Angus ZG, Rozen WM. Advancements in microsurgery: a comprehensive systematic review of artificial intelligence applications. *J Plast Reconstr Aesthet Surg.* 2025;101:65-76.
- Colceriu AA, Ifteni PI, Rogozea L, Leășu FG, Stuckler D. The impact of artificial intelligence on facial aesthetic surgery: a systematic review. *Aesthet Surg J Open Forum.* 2026;8:ojag114.
- Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378.
- Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiol Artif Intell.* 2024;6(4):e240300.
- Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ.* 2025;388:e081554.
- Chang CW, Lai F, Christian M, Chen YC, Hsu C, Chen YS, et al. Deep learning-assisted burn wound diagnosis: diagnostic model development study. *JMIR Med Inform.* 2021;9(12):e22798.
- Chang CW, Ho CY, Lai F, Christian M, Huang SC, Chang DH, et al. Application of multiple deep learning models for automatic burn wound assessment. *Burns.* 2023;49(5):1039-1051.
- El Masry M, Rahman MM, Gnyawali SC, Smeenge T, Xue Y, Hartman BC, et al. Automated noninvasive burn diagnostic system for health care using artificial intelligence: AMBUSH-AI. *Ann Surg.* 2026;284(2):296-304.
- Zhang BH, Chen K, Lu SM, Nakfoor B, Cheng R, Gibstein A, et al. Turning back the clock: artificial intelligence recognition of age reduction after face-lift surgery correlates with patient satisfaction. *Plast Reconstr Surg.* 2021;148(1):45-54.
- Lo LJ, Yang CT, Ho CT, Liao CH, Lin HH. Automatic assessment of 3-dimensional facial soft tissue symmetry before and after orthognathic surgery using a machine learning model: a preliminary experience. *Ann Plast Surg.* 2021;86(3 Suppl 2):S224-8.
- Shin W, Yeom HG, Lee GH, Yun JP, Jeong SH, Lee JH, et al. Deep learning based prediction of necessity for orthognathic surgery of skeletal malocclusion using cephalogram in Korean individuals. *BMC Oral Health.* 2021;21(1):130.
- de Oliveira PHJ, Li T, Li H, Gonçalves JR, Santos-Pinto A, Gandini Junior LG, et al. Artificial intelligence as a prediction tool for orthognathic surgery assessment. *Orthod Craniofac Res.* 2024;27(5):785-794. doi:10.1111/ocr.12805. [Verify complete bibliographic record.]
- Ho CT, Lo LJ, Chiang WC, Liu CM, Lin HH. Quantification of facial symmetry in orthognathic surgery: a novel approach integrating 3D contour maps and hyper-dimensional computing. *Comput Biol Med.* 2024;183:109189.
- Alper DP, Almeida MN, Hosseini H, De Baun HM, Moscarelli J, Hu KG, et al. Perceived age and gender perception using facial recognition software following facial feminization surgery. *J Craniofac Surg.* 2024;35(1):39-42.
- Du H, Liang H, Peng B, Qi Z, Jin X. Age reduction after face-lift surgery in Chinese population: an outcome study using artificial intelligence and objective observer-based assessment. *Aesthetic Plast Surg.* 2024;48(23):4760-8.
- Asaad M, Lu SC, Hassan AM, Kambhampati P, Mitchell D, Chang EI, et al. The use of machine learning for predicting complications of free-flap head

- and neck reconstruction. *Ann Surg Oncol.* 2023;30(4):2343-52.
21. Huang RW, Tsai TY, Hsieh YH, Hsu CC, Chen SH, Lee CH, et al. Reliability of postoperative free flap monitoring with a novel prediction model based on supervised machine learning. *Plast Reconstr Surg.* 2023;152(5):943e-52e.
 22. Kim J, Lee SM, Kim DE, Kim S, Chung MJ, Kim Z, et al. Development of an automated free flap monitoring system based on artificial intelligence. *JAMA Netw Open.* 2024;7(7):e2424299.
 23. Kim DK, Corpuz GS, Ta CN, Weng C, Rohde CH. Applying unsupervised machine learning approaches to characterize autologous breast reconstruction patient subgroups: an NSQIP analysis of 14,274 patients. *J Plast Reconstr Aesthet Surg.* 2024;88:330-9.
 24. Chen Y, Zhang A, Wang ML, Black GG, Zhou G, Otterburn DM. Using a machine learning approach to predict the need for elective revision and unplanned surgery after implant-based breast reconstruction. *Plast Reconstr Surg Glob Open.* 2024;12(3):e5542.
 25. Ozmen BB, Phuyal D, Berber I, Schwarz GS. Artificial intelligence prediction model for readmission after DIEP flap breast reconstruction based on NSQIP data. *J Plast Reconstr Aesthet Surg.* 2025;106:1-8.
 26. Gabay AJ, Chen J, Stern CS, Gibbons C, Mehrara B, Nelson JA. Predicting breast reconstruction readmission, reoperation, and prolonged length of stay: a machine learning approach. *J Surg Oncol.* 2025;132(2):345-53.
 27. Chen J, Gabay AJ, Boe LA, Shammass RL, Stern CS, Pusic A, et al. Machine learning accurately predicts patient-reported outcomes 1 year after breast reconstruction. *Ann Surg.* 2026;284(1):176-83.
 28. Hung YC, Chaker SC, Sigel M, Saad M, Slater ED. Comparison of patient education materials generated by Chat Generative Pre-Trained Transformer versus experts: an innovative way to increase readability of patient education materials. *Ann Plast Surg.* 2023;91(4):409-12.
 29. Abi-Rafeh J, Mroueh VJ, Bassiri-Tehrani B, Marks J, Kazan R, Nahai F. Complications following body contouring: performance validation of Bard, a novel AI large language model, in triaging and managing postoperative patient concerns. *Aesthetic Plast Surg.* 2024;48(5):953-76.
 30. Ozmen BB, Berber I, Schwarz GS. Initial proof-of-concept study for a plastic surgery-specific artificial intelligence large language model: PlasticSurgeryGPT. *Aesthet Surg J.* 2025;45(8):860-4.
 31. Perez Rivera LR, Gursky AK, Elmer N, Boyd CJ, Karp NS. Evaluating the quality and reliability of large language models for plastic surgery patient education: a comparative analysis of ChatGPT and OpenEvidence. *Aesthet Surg J.* 2026;46(2):160-167.
 32. U.S. Food and Drug Administration. Good machine learning practice for medical device development: guiding principles. Silver Spring (MD): U.S. Food and Drug Administration. Available from: <https://www.fda.gov/medical-devices/artificial-intelligence-enabled-medical-devices/good-machine-learning-practice-medical-device-development-guiding-principles> Accessed on 28 August 2026.

Cite this article as: Fuentes SMS, Rosales MAC, Cisneros ANG, Garza MSR, González LFH, Márquez VG, et al. Artificial intelligence in plastic and reconstructive surgery: clinical applications, translational readiness and barriers to real-world implementation. *Int J Res Med Sci* 2026;14:5460-8.