ISSN: 2688-8645

Foundation Models for Medical Image Segmentation: Technical Progress, Clinical Translation, and Governance Challenges

Juntao Wei *

School of Health and Medical Technology, Chengdu Neusoft University, Chengdu, Sichuan, 611844, China

* Correspondence author Email: 2957364608@qq.com

Abstract

Medical image segmentation is moving from task-specific convolutional models toward foundation models that can be adapted across organs, modalities, and clinical tasks with fewer manual labels. This transition has been accelerated by self-supervised pretraining, vision-language learning, and promptable segmentation frameworks such as the Segment Anything Model and its medical derivatives. However, the clinical value of these systems cannot be inferred from technical novelty alone. Medical images differ from natural images in dimensionality, intensity statistics, acquisition protocols, disease prevalence, and safety requirements, and recent evaluations show that naive zero-shot transfer remains inconsistent across modalities and lesion types. This narrative review synthesizes literature published up to May 22, 2026, on foundation models for medical image segmentation, with emphasis on technical evolution, application scenarios, validation strategies, and governance needs. Current evidence suggests that foundation models are most promising when they are deployed as interactive, auditable components within human-in-the-loop workflows, where they can reduce annotation burden, support rapid draft segmentation, and improve consistency across large imaging studies. Their translation into routine practice requires external validation, uncertainty-aware quality control, prospective workflow evaluation, bias assessment, and transparent reporting under medical AI guidelines. Future work should prioritize patient-level multimodal modeling, 3D and longitudinal segmentation, federated evaluation, and clinically meaningful endpoints rather than isolated benchmark gains.

Keywords

Medical Image Segmentation; Foundation Models; Segment Anything Model; Deep Learning; Multimodal Imaging; Clinical Translation; Artificial Intelligence.

1. Introduction

Segmentation is one of the central technical operations in medical imaging. It converts images into organ boundaries, lesion masks, tissue compartments, and quantitative biomarkers that can support diagnosis, radiotherapy planning, surgical navigation, longitudinal follow-up, and population-scale imaging research. Deep learning transformed this task by replacing hand-crafted features with hierarchical representations learned from annotated data, and early reviews already showed that medical image analysis had become one of the most active clinical domains for convolutional neural networks. [1] U-Net provided a practical encoder-decoder architecture for biomedical segmentation under limited labels, while nnU-Net later demonstrated that careful self-configuration of preprocessing, architecture, training schedule, and post-processing can match or outperform many bespoke methods across heterogeneous segmentation tasks.[2, 3] The Medical Segmentation Decathlon further made clear that robust

ISSN: 2688-8645

segmentation requires more than a strong network: it requires cross-organ, cross-modality, and cross-institutional evaluation. [4]

Despite these advances, conventional supervised segmentation remains constrained by annotation cost, expert scarcity, task-specific retraining, and poor transfer when scanners, protocols, disease patterns, or patient populations shift. Three-dimensional CT and MRI segmentation requires voxel-level labeling that is expensive and often ambiguous at organ boundaries. Lesion segmentation is even harder because pathologic structures may be small, low contrast, anatomically variable, or underrepresented in training data. These limitations have encouraged a shift toward reusable representation learning. Self-supervised medical models such as Models Genesis showed that useful 3D representations could be learned from imaging data before task-specific fine-tuning.[5] More broadly, foundation models have been proposed as general-purpose systems trained on broad data and adapted to downstream clinical tasks through prompting, fine-tuning, or multimodal conditioning. [6]

The recent rise of promptable segmentation models has made this transition visible to clinicians and imaging researchers. The Segment Anything Model (SAM) introduced a general promptable segmentation paradigm for natural images, enabling mask generation from points, boxes, and other prompts.[7] In medical imaging, MedSAM and related systems adapted this idea to large-scale medical segmentation, while independent evaluations showed that zero-shot SAM performance can be highly variable when applied directly to CT, MRI, ultrasound, microscopy, endoscopy, and pathology images. [8-11] The field is therefore entering a useful but delicate phase. Foundation models may reduce the marginal cost of segmentation, but they also risk creating a false sense of generality if benchmark performance is separated from clinical context. This review argues that the relevant question is not whether foundation models can replace established segmentation pipelines, but how they can be safely integrated into imaging workflows that demand accuracy, transparency, accountability, and clinical usefulness.

2. Search Strategy and Review Scope

This narrative review was developed through targeted searches of PubMed and Crossref up to May 22, 2026. Search terms combined "medical image segmentation", "foundation model", "Segment Anything", "vision-language model", "self-supervised learning", "clinical translation", "medical imaging", "reporting guideline", and "trustworthy artificial intelligence". Priority was given to peer-reviewed articles, major conference papers, and reporting or governance guidelines with verifiable DOI or arXiv metadata. The review is not a systematic review or meta-analysis; it does not pool performance metrics across studies because model architectures, prompting protocols, datasets, imaging modalities, and endpoints remain too heterogeneous for a reliable pooled estimate. Instead, it synthesizes the technical and translational patterns most relevant to authors preparing segmentation models for clinical or journal evaluation.

3. Technical Evolution of Medical Image Segmentation

The first phase of modern medical image segmentation was dominated by supervised convolutional architectures. U-Net remains influential because its skip-connected encoder-decoder design preserves localization while learning high-level context, a combination well suited to biomedical images with limited training samples. [2] nnU-Net then shifted attention from isolated architecture novelty to pipeline reliability. Its success showed that many apparent algorithmic improvements disappear when preprocessing, patch size, augmentation, inference, and post-processing are tuned fairly across datasets. [3] This lesson remains important in the foundation model era: a large pretrained model is not automatically superior to a well-calibrated domain-specific baseline.

ISSN: 2688-8645

Benchmark datasets changed the field by revealing how task diversity affects segmentation. The Medical Segmentation Decathlon included multiple organs, pathologies, and modalities, emphasizing that generalization must be tested across anatomical and acquisition variation rather than inferred from a single dataset.[4] TotalSegmentator extended this practical agenda by providing robust CT segmentation of many anatomic structures, illustrating how broadly useful segmentation tools can support downstream radiology and research workflows when trained and evaluated at scale. [12] These systems are not foundation models in the broadest sense, but they set the performance and usability standards that foundation models must meet or exceed.

Self-supervised and multimodal pretraining created a bridge between supervised segmentation and foundation models. Models Genesis demonstrated that generic 3D image representations can be learned from medical volumes and transferred to downstream tasks.[5] In chest radiography, self-supervised vision-language learning enabled expert-level detection of pathologies from image-report pairs without manual disease labels, illustrating the value of pairing imaging data with routinely generated clinical text.[19] Generalist medical AI and medical foundation model work has since argued that large models can become adaptable substrates for classification, segmentation, report generation, retrieval, and decision support if they are trained on sufficiently diverse clinical data and evaluated under clinically meaningful conditions. [6, 21, 22] For segmentation, this means that the model should not only produce masks but also interact with prompts, reports, prior studies, uncertainty estimates, and user corrections.

4. Promptable Segmentation and the Medical SAM Family

SAM changed the segmentation interface by separating image encoding from prompt-conditioned mask decoding. In natural images, this design allows the user to guide segmentation with points or boxes while the model uses a broadly trained visual representation to generate masks. [7] Medical imaging poses a harder transfer problem. CT, MRI, ultrasound, endoscopy, microscopy, and pathology images differ from natural photographs in texture, contrast mechanisms, dimensionality, and object definition. Boundaries may be physiologic rather than visually sharp, lesions may be defined by clinical context, and the same anatomical structure may appear differently across acquisition protocols. Direct zero-shot use of SAM therefore produces mixed results: some large structures can be segmented acceptably with well-placed prompts, while small lesions, low-contrast tissues, volumetric structures, and modality-specific artifacts often expose failure modes.[9, 10]

Domain adaptation has been the main response. MedSAM trained SAM-style segmentation on a large collection of medical images and showed that medical-domain adaptation improves generality across many segmentation tasks. [8] LeSAM focused on lesion segmentation, where ambiguity, small target size, and heterogeneous appearance make prompt design especially important.[13] FastSAM3D and prompt-driven 3D segmentation studies explored volumetric adaptation, a necessary step for radiology because clinical CT and MRI decisions are made across slices rather than on isolated 2D frames.[14, 15] Segment Anything for Microscopy extended the promptable paradigm into microscopy, where object density and instance boundaries differ sharply from radiology.[16] Emerging 2026 studies on efficient and zero-shot medical segmentation indicate that the field is now optimizing speed, annotation efficiency, and zero-shot robustness rather than simply proving that SAM-like systems can be adapted to medicine. [17, 18]

The practical value of promptable medical segmentation should be evaluated at the workflow level. A model that achieves high Dice similarity with repeated expert prompting may still be inefficient if the user must provide many corrective prompts. Conversely, a model with modest

ISSN: 2688-8645

fully automatic performance may be clinically useful if it provides a fast editable draft that reduces contouring time without increasing failure risk. Useful evaluations should therefore report prompt type, number of prompts, user expertise, edit time, surface distance, small-object performance, uncertainty or quality flags, and external validation across scanners and institutions. This evaluation philosophy aligns with medical AI reporting standards that ask authors to describe data sources, preprocessing, model development, validation, intended use, and human-AI interaction rather than reporting a single headline score. [30-32]

5. Application Scenarios in Medical Imaging

Radiology is the most direct application setting for foundation segmentation models because CT and MRI produce large volumetric studies where manual segmentation is time-consuming. Organ segmentation can support opportunistic screening, radiotherapy planning, surgical navigation, and automated measurement. TotalSegmentator shows how multi-organ CT segmentation can become infrastructure for downstream analysis when it is packaged as a robust tool rather than a one-off algorithm. [12] Foundation models could extend this idea to new organs, rare variants, and local protocols with fewer labels. Tumor and lesion segmentation is harder because boundaries often depend on contrast phase, prior imaging, radiologist judgment, and treatment context. Domain-adapted systems such as LeSAM [13] are promising, but clinical adoption still requires use-case-specific evaluation.

Digital pathology, microscopy, ophthalmology, and echocardiography broaden the concept of medical imaging foundation models. Pathology foundation models have been trained on very large slide collections and can support tissue representation learning, retrieval, weakly supervised classification, and multimodal report-image tasks. [20, 21] Vision-language models for biomedical tasks and echocardiogram interpretation show that models can connect visual patterns to clinical language, which is important because many imaging tasks require reasoning over both pixels and reports. [22, 23] Precision oncology foundation models further suggest that imaging, pathology, and molecular context may eventually be combined within patient-specific models. [24] For segmentation, these developments matter because the boundary of a lesion or tissue compartment is often not purely visual; it is influenced by diagnosis, staging, histology, and therapeutic intent.

Human-in-the-loop use is likely to be the most realistic near-term pathway. In radiotherapy, pathology, and research annotation, segmentation models can produce first-pass masks that experts correct. In clinical radiology, an interactive tool may be safer than a fully automatic black-box output because it keeps the clinician in control while still reducing repetitive labor. However, human-in-the-loop deployment is not automatically safe. Automation bias, time pressure, and poor interface design can lead users to accept flawed masks. Clinical studies should therefore measure not only segmentation accuracy but also editing behavior, time saved, error detection, user trust calibration, and downstream decision impact.[28, 29, 35]

6. Clinical Translation: Validation, Bias, and Governance

The central translational challenge is generalization. Medical AI systems can perform well on internal test sets and fail after deployment because patient mix, disease prevalence, scanner vendor, acquisition protocol, reconstruction kernel, annotation culture, and clinical workflow differ across sites. A pneumonia detection study showed variable generalization across institutions, and fairness studies in medical imaging have shown that demographic imbalance can produce biased classifiers and underdiagnosis in underserved groups. [25-27] Segmentation models are exposed to similar risks. A model may trace organ boundaries well in common anatomy but fail in pediatric patients, postoperative anatomy, rare tumors, implanted devices, motion artifacts, or low-resource settings underrepresented during training.

ISSN: 2688-8645

Foundation models add new governance questions. Because these models are pretrained on broad datasets, developers must document data provenance, de-identification, modality coverage, population characteristics, annotation sources, and known exclusions. If a model is adapted after deployment, version control and post-market monitoring become essential because silent updates can change behavior. For clinical segmentation, uncertainty-aware quality control is especially important. A system should be able to flag masks that are out of distribution, internally inconsistent, or likely to need expert review. Without such safeguards, foundation models may scale errors as efficiently as they scale performance.

Reporting guidelines provide a practical route to better evidence. CLAIM and its 2024 update specify what medical imaging AI studies should report so that readers can evaluate data, model development, validation, and reproducibility. [30, 31] MI-CLAIM addresses minimum information for clinical AI modeling, while CONSORT-AI, SPIRIT-AI, DECIDE-AI, TRIPOD+AI, and FUTURE-AI extend reporting and evaluation expectations across trials, protocols, early-stage clinical evaluation, prediction models, and trustworthy deployment. [32-37] A foundation segmentation paper prepared for journal submission should therefore include a clear intended use, transparent dataset description, external validation, baseline comparison against strong non-foundation models, prompt and user-interaction protocol, subgroup analysis, failure examples, and a deployment plan proportionate to clinical risk.

7. Future Directions

First, segmentation foundation models need stronger 3D and longitudinal reasoning. Many current promptable systems inherit a 2D image bias, yet radiology decisions frequently depend on volumetric continuity, temporal change, and spatial relationships across organs. Future models should integrate slice context, prior examinations, scanner metadata, and anatomical constraints so that segmentation is stable across the full study rather than locally plausible on selected slices. Transformer-based segmentation research and recent reviews point toward architectures that can integrate global context, but clinical usefulness will depend on whether these models remain efficient and interpretable in routine workflows. [38]

Second, multimodal conditioning should become clinically grounded rather than merely larger. The strongest future systems will use radiology reports, pathology notes, laboratory data, prior masks, and treatment intent to guide segmentation when visual boundaries are ambiguous. This does not mean that text should override image evidence. Instead, clinical context should help define the target, select the relevant phase or sequence, and determine which uncertainty requires expert review. Vision-language foundation models in pathology, echocardiography, and precision oncology suggest that such integration is technically feasible, but segmentation-specific validation remains necessary.[20, 23, 24]

Third, the field needs benchmark designs that reward reliability, not only peak accuracy. Public leaderboards should include external sites, rare subgroups, scanner shifts, prompt-effort measures, edit-time measures, failure detection, and prospective workflow endpoints. Federated and privacy-preserving evaluation may allow hospitals to test models without centralizing sensitive data. Journals can accelerate this shift by requiring reference baselines, standardized reporting, and release of inference code or sufficient implementation detail. The foundation model era will be clinically meaningful only if models become easier to test, audit, and correct.

8. Limitations

This review has limitations. It is a narrative synthesis rather than a systematic review, and it does not provide a pooled estimate of segmentation accuracy. The literature is changing rapidly, with many important models first appearing as preprints, software releases, or challenge

submissions before formal peer review. To reduce citation risk, this manuscript prioritized peer-reviewed or otherwise verifiable sources, which may underrepresent very recent unpublished systems. The paper also emphasizes segmentation and adjacent vision-language imaging models; it does not comprehensively review classification, reconstruction, registration, or report generation. Finally, clinical translation requirements vary by country, specialty, and risk category, so the governance recommendations should be adapted to the target jurisdiction and intended use.

9. Conclusion

Foundation models are reshaping medical image segmentation by changing both the technical pipeline and the user interface. The most important advance is not simply larger pretraining; it is the possibility of adaptable segmentation systems that can respond to prompts, reuse broad representations, and fit into expert correction workflows. Current evidence supports cautious optimism. Domain-adapted models such as MedSAM and related systems show that promptable segmentation can be made more suitable for medical images, while evaluations of zero-shot transfer remind the field that medical reliability cannot be assumed from natural-image success. For journal-ready research and clinical translation, foundation segmentation models should be evaluated against strong baselines, tested externally, reported transparently, audited for bias, and assessed in workflows where expert users remain able to detect and correct errors. If future work prioritizes these standards, foundation models may become practical infrastructure for precise, efficient, and accountable medical imaging.

Data Availability Statement

No new datasets were generated or analyzed in this narrative review. The literature search and DOI verification records are available from the corresponding author on reasonable request.

Ethics Declaration

This article does not report original studies involving human participants, human tissue, animal subjects, or identifiable private data. Formal ethics approval was therefore not required.

Competing Interests

The authors declare no competing interests.

References

[1] Litjens, G., Kooi, T., Ehteshami Bejnordi, B., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. https:// doi. org/ 10.1016/j.media.2017.07.005.

[2] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

[3] Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z.

[4] Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., et al. (2022). The Medical Segmentation Decathlon. *Nature Communications*, 13(1), 4128. https://doi.org/ 10. 1038/ s41467-022-30695-9.

[5] Zhou, Z., Sodha, V., Pang, J., Gotway, M. B., & Liang, J. (2021). Models Genesis. *Medical Image Analysis*, 67, 101840. https://doi.org/10.1016/j.media.2020.101840.

[6] Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., et al. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956), 259–265. https: // doi.org/10.1038/s41586-023-05881-4.

[7] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment Anything [Preprint]. arXiv. https://arxiv.org/abs/2304.02643

[8] Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B., et al. (2024). Segment anything in medical images. *Nature Communications*, 15(1), 654. https://doi.org/10.1038/s41467-024-44824-z.

[9] Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., et al. (2024). Segment anything model for medical images? *Medical Image Analysis*, 92, 103061. https://doi.org/ 10. 1016/ j. media. 2023. 103061.

[10] Mazurowski, M. A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y., et al. (2023). Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis*, 89, 102918. https: // doi. org/10.1016/j.media.2023.102918.

[11] Zhang, Y., Shen, Z., & Jiao, R. (2024). Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, 171, 108238. https://doi.org/10.1016/j.compbiomed.2024.108238.

[12] Wasserthal, J., Breit, H. C., Meyer, M. T., Pradella, M., Hinck, D., Sauter, A. W., et al. (2023). TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence*, 5(5), e230024. https://doi.org/10.1148/ryai.230024.

[13] Gu, Y., Wu, Q., Tang, H., Mai, X., Shu, H., Li, B., et al. (2024). LeSAM: Adapt Segment Anything Model for Medical Lesion Segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(10), 6031–6041. https://doi.org/10.1109/JBHI.2024.3406871.

[14] Shen, Y., Li, J., Shao, X., Romillo, B. I., Jindal, A., Dreizin, D., et al. (2024). FastSAM3D: An Efficient Segment Anything Model for 3D Volumetric Medical Images. *Lecture Notes in Computer Science*, 542–552. https://doi.org/10.1007/978-3-031-72390-2_51.

[15] Li, H., Liu, H., Hu, D., Wang, J., & Oguz, I. (2024). PROMISE: Prompt-Driven 3D Medical Image Segmentation Using Pretrained Image Foundation Models. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)* (pp. 1–5). https://doi.org/10.1109/ISBI56570.2024.10635207.

[16] Archit, A., Freckmann, L., Nair, S., Khalid, N., Hilt, P., Rajashekar, V., et al. (2025). Segment Anything for Microscopy. *Nature Methods*, 22(3), 579–591. https://doi.org/10.1038/s41592-024-02580-4.

[17] Wu, X. T., Chen, X. D., Wu, W., Ma, W., & Song, H. (2026). Accelerating vision foundation model for efficient medical image segmentation. *Medical Physics*, 53(1). https://doi.org/10.1002/mp.70193.

[18] Zhang, R., Huang, M., & Li, R. (2026). MedZeroSeg: Zero-shot medical image segmentation via vision foundation models. *PLOS ONE*, 21(3), e0344978. https://doi.org/10.1371/journal.pone.0344978.

[19] Tiu, E., Talius, E., Patel, P., Langlotz, C. P., Ng, A. Y., Rajpurkar, P., et al. (2022). Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12), 1399–1406. https://doi.org/10.1038/s41551-022-00936-9.

[20] Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Liang, I., Ding, T., et al. (2024). A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3), 863–874. https:// doi. org/ 10. 1038/s41591-024-02856-4.

[21] Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., et al. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3), 850–862. https://doi.org/10.1038/s41591-024-02857-3.

[22] Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., et al. (2024). A generalist vision-language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11), 3129–3141. https:// doi.org/10.1038/s41591-024-03185-2.

[23] Christensen, M., Vukadinovic, M., Yuan, N., & Ouyang, D. (2024). Vision-language foundation model for echocardiogram interpretation. *Nature Medicine*, 30(5), 1481–1488. https://doi.org/ 10. 1038/ s41591-024-02959-y.

[24] Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., et al. (2025). A vision-language foundation model for precision oncology. *Nature*, 638(8051), 769–778. https://doi.org/10.1038/s41586-024-08378-w.

[25] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., Oermann, E. K., et al. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11), e1002683. https://doi.org/ 10.1371/ journal. pmed. 1002683.

[26] Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23), 12592–12594. https://doi.org/ 10.1073/ pnas. 1919 012117.

[27] Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12), 2176–2182. https://doi.org/ 10. 1038/ s4 1591-021-01595-0.

[28] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1), 195. https://doi.org/ 10. 1186/ s12916-019-1426-2.

[29] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0.

[30] Mongan, J., Moy, L., & Kahn, C. E. Jr. (2020). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiology: Artificial Intelligence*, 2(2), e200029. https://doi.org/10.1148/ryai.2020200029.

[31] Tejani, A. S., Klontzas, M. E., Gatti, A. A., Mongan, J. T., Moy, L., Park, S. H., et al. (2024). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiology: Artificial Intelligence*, 6(4), e240300. https://doi.org/10.1148/ryai.240300.

[32] Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., et al. (2020). Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature Medicine*, 26(9), 1320–1324. https://doi.org/10.1038/s41591-020-1041-y.

[33] Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ*, 370, m3164. https://doi.org/10.1136/bmj.m3164.

[34] Cruz Rivera, S., Liu, X., Chan, A. W., Denniston, A. K., & Calvert, M. J. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ*, 370, m3210. https://doi.org/10.1136/bmj.m3210.

[35] Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., et al. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, 28(5), 924–933. https://doi.org/10.1038/s41591-022-01772-9.

[36] Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., et al. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378. https://doi.org/10.1136/bmj-2023-078378.

[37] Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., et al. (2025). FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, 388, e081554. https://doi.org/10.1136/bmj-2024-081554

[38] Xu, Y., Peng, Y., Zhang, C., Jiang, K., She, X., Feng, L., et al. (2026). Application of transformer models in medical image segmentation: a narrative review. *Quantitative Imaging in Medicine and Surgery*, 16(5), 421. https://doi.org/10.21037/qims-2025-aw-2381.