Academia Open

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

Academia Open

By Universitas Muhammadiyah Sidoarjo

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

Table Of Contents

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

Originality Statement

The author[s] declare that this article is their own work and to the best of their knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the published of any other published materials, except where due acknowledgement is made in the article. Any contribution made to the research by others, with whom author[s] have work, is explicitly acknowledged in the article.

Conflict of Interest Statement

The author[s] declare that this article was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright Statement

Academia Open

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

EDITORIAL TEAM

Academia Open

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

Article information

Check this article update (crossmark)

Check this article impact (*)

Save this article to Mendeley

(*) Time for indexing process is various, depends on indexing database platform

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

Foundation Models in Medical Imaging: A Systematic Review, Taxonomy, and Future Research Agenda

Roaa Ghanim , roaa.ghanim11@ec.edu.iq (*)

General Directorate of Education in Al- Qadisiyah Govemorate, Diwaniyah, Iraq

(*) Corresponding author

Abstract

Medical artificial intelligence relies heavily on imaging for lesion detection and anatomical quantification, providing essential support for cancer diagnosis and treatment planning (**General Background**). While foundation models are rapidly transforming radiological workflows, traditional evaluations focus predominantly on technical segmentation metrics rather than clinical utility (**Specific Background**). Existing literature lacks a comprehensive synthesis combining model taxonomies, multi-task assessments beyond segmentation, and standardized clinical readiness appraisals (**Knowledge Gap**). This systematic review evaluated forty-two studies published between 2024 and 2026 across eleven databases using PRISMA 2020 guidelines and AI-specific quality frameworks (**Aims**). Results demonstrated high organ segmentation accuracy with Dice coefficients exceeding 0.90, whereas performance declined for infiltrative tumours, and external validation was present in only one-third of studies (**Results**). We propose an eight-level clinical-readiness scale and research roadmap to transition medical AI from technical benchmarks to deployment (**Novelty**). Standardized validation protocols are critical for advancing safe decision support (**Implications**).

Key Findings Highlights

Current medical foundation models achieve high technical precision in organ segmentation but show reduced performance in low-contrast and post-treatment lesions.

Less than one-third of evaluated studies report external validation, and no interventional prospective trials currently exist.

Transitioning AI to routine oncology care requires establishing standardized clinical endpoints, fairness reporting, and multi-institutional data frameworks.

Keywords: Foundation Models, Medical Imaging, Tumour Quantification, Clinical Decision Support, Systematic Review

Published date: 2026-09-09

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

1. Introduction

1.1 Background: imaging in organ and tumour care

Today, medical imaging plays a key role in the treatment of organ diseases and cancer. Computed tomography (CT), magnetic resonance imaging (MRI), positron-emission tomography (PET), ultrasound, radiography, and digital pathology, collectively, aid in the diagnosis, staging, treatment planning, intra-procedural guidance, and longitudinal follow-up. In oncology, imaging determines the anatomical and biological extent of disease, is the basis for target delineation for surgery or radiotherapy, and serves as the measure to assess response according to criteria like RECIST 1.1 (Eisenhauer et al., 2009). As imaging volumes increase, these decisions rely on accurate and reproducible quantification of organs and lesions, which is labour intensive, inter-reader variable and hard to maintain. The automation, generalization and analysis of images is therefore not a luxury but a necessity for the scalable and equitable provision of cancer care.

The clinical and operational risk is high. Cancer is one of the top killers around the world and imaging plays a vital role in the patient journey, from incidental detection and characterisation to the staging, choice of therapy and monitoring for recurrence. However, the same success of imaging has come with a problem for sustainability: the number of scans and the complexity of the images has increased more rapidly than the number of radiologists and pathologists, adding to the length of turnaround time and the burden on the reporting process. Currently, manual delineation of organs and tumours is the gold standard for many quantitative tasks; this is time-consuming and non-trivially variable between and within observers, leading to propagation of this variation throughout the downstream process of making decisions, e.g. radiotherapy dose planning and response classification. It is increasingly recognised that the ability to provide automated analysis that is reproducible and scalable is a structural need for cancer care, especially when specialist expertise is limited and is therefore increasingly considered as a necessary component of equitable and timely cancer care.

1.2 Classical Segmentation to Foundation Models

In the majority of the deep-learning era, medical image analysis systems have been largely comprised of task-specific convolutional and transformer networks. This paradigm was best captured by the self-configuring nnU-Net framework, which could set up a single pipeline for many tasks, achieving state-of-the-art performance for many benchmarks (Isensee et al., 2021). Such specialist models are only suitable for the distribution of the training data, but are not transferable between modalities, scanners, institutions and disease types, and each new task often requires a new annotation and retraining.

This limitation is starting to be broken by the introduction of large, widely trained foundation models (Azad et al., 2023; He et al., 2024). Promptable, zero-shot segmentation of natural images was achieved with Segment Anything Model (SAM) (Kirillov et al., 2023), while its medical extension MedSAM achieved impressive performance on 86 internal and 60 external tasks against modality-specific specialists on over 1.5 million image–mask pairs across ten modalities and over thirty cancer types (Ma et al., 2024). Volumetric models were developed along the same lines as SAM-Med3D (Wang et al., 2025) and SegVol (Du et al., 2024); large multi-structure tools were created like TotalSegmentator (Wasserthal et al., 2023); self-supervised pathology encoders such as UNI (Chen et al., 2024); vision-language and multimodal systems like CONCH (Lu et al., 2024), CheXagent (Chen, Z., et al., 2024), BiomedGPT (Zhang, K., et al., 2024); and generalist biomedical models (Moor et al., 2023a; Tu et al., 2024). These systems as a whole march in the direction of a generic medical AI that can handle heterogenous data without requiring much task-specific supervision (Moor et al., 2023a).

Such systems can vary in their architecture in addition to the way they're adjusted to a clinical task; in addition, the type of adaptation strategy has a significant impact on performance, data requirement, and safety. Zero-shot use: A pre-trained model is used without further training.Few-shot use: A few labelled examples are given to a pre-trained model. Promptable models like MedSAM represent interactive adaptation with human-in-the-loop while self-supervised encoders are usually frozen and probed. This review extracts adaptation strategy as a primary descriptor, and uses this strategy as a part of any assessment of clinical readiness, rather than a detail of implementation because each strategy has different generalisation and failure traits.

Beyond accurately defining boundaries, it is important to assert exactly what ‘organ and tumour understanding’ entails. In the clinical context, understanding involves establishing the presence or absence of a structure, quantifying its size, volume and time course, correlating imaging phenotype with biology, prognosis and probable treatment response, and communicating this in a way that aids in a decision. The value of imaging is added by these latter steps, which involve bringing in context, prior studies and in many cases non-imaging data, but these latter steps are crucial to the diagnostic and prognostic value of imaging. This is why foundation and multimodal models are appealing from a theoretical perspective, a single pre-trained backbone can be used along this entire chain how well they can perform these models is what this review aims to uncover.

1.3 Problem statement

Nevertheless, literature reports the quality of the foundation models mainly based on technical metrics that are mostly related to the segmentation by looking at the Dice similarity coefficient. The value of high segmentation accuracy remains mostly unproven in terms of organ/tumor understanding, quantification of tumor burden, estimation of tumor response to therapy and prognosis, and trustworthy decision support. A model can be very accurate in Dice but fail to produce accurate measurements of a meaningful change in lesion volume, fail to be robust in the way it performs across different hospitals, or write fluent but unsafe text. The overall issue is then one of ‘measure of the wrong thing’.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

1.4 Literature gap

The present reviews are beneficial, but incomplete. Some point to the lack of exploration of real-world applications and only limited testing on unseen data (Mazurowski et al., 2023; Noh & Lee, 2025; Huang, Y., et al., 2024; Zhang, Y., et al., 2024). No review has combined (i) the model families with organs, modalities, adaptation strategies, and clinical-readiness levels in a single taxonomy, (ii) AI-specific reporting and risk-of-bias frameworks to appraise the 2024–2026 evidence, and (iii) been both a review and extended beyond segmentation to quantification, response assessment, prognosis, report generation, and decision support. The evidence is growing quickly but is still disjointed and many studies do not include external validation, prospective analysis, fairness analysis, uncertainty analysis, and workflow testing.

1.5 Objectives

Primary objective. To systematically review studies published between 2024 and 2026 on foundation and large medical AI models for organ and tumour understanding, quantification, and clinical decision support.

Secondary objectives:

- To classify existing models using a clinically meaningful, six-family taxonomy.
- To summarise applications across organs, tumours, imaging modalities, and clinical tasks.
- To evaluate validation quality and to grade clinical readiness on a transparent scale.
- To identify limitations, evidence gaps, and prioritised directions for future research.

2. Methods

2.1 Protocol and registration

The search for the review was designed beforehand and is reported in line with PRISMA 2020 (Page et al., 2021). The protocol was prepared and defined prior to screening, including the research questions, eligibility criteria, search strategy, data-extraction fields, appraisal tools, and a synthesis plan. It is recommended that prospective registration of executed protocol be provided in PROSPERO or the Open Science Framework and that the registration number, date and any amendments be noted in the final version of the manuscript. After screening began, there were no changes in the methods, which included the eligibility criteria and primary outcomes.

2.2 Reporting guideline and evaluation frameworks

The reporting is based on the PRISMA 2020 statement and flow diagram (Page et al., 2021) and the TRIPOD+AI statement where prediction-model reporting was relevant (Collins et al., 2024). Since these were AI studies and not traditional diagnostic or therapeutic studies, we used AI-specific frameworks matched to study type: Checklist for Artificial Intelligence in Medical Imaging (CLAIM) 2024, reporting completeness (Tejani et al., 2024); PROBAST+AI, prediction and prognostic models (Moons et al., 2025); QUADAS-2, diagnostic-accuracy studies (Whiting et al., 2011); and DECIDE-AI, early-stage clinical evaluation of decision-support systems (Vasey et al., 2022). We extracted items of fairness, robustness and transparency from the dimensions of trustworthiness in FUTURE-AI (Lekadir et al., 2025). CONSORT-AI and SPIRIT-AI, along with the Cochrane risk-of-bias tool for randomised trials (RoB 2) are only applicable to randomised trials and trial protocols (Liu et al., 2020; Cruz Rivera et al., 2020), however, no randomised trials met our criteria, and this absence is itself a significant finding (Section 3), which is why these instruments were retained in the protocol, but not applied.

2.3 Research questions

1. Which foundation and large medical AI models have been applied to organ and tumour analysis during 2024–2026?
2. Which clinical tasks beyond segmentation are addressed?
3. Which organs, tumours, and imaging modalities are most studied?
4. Which adaptation strategies are used (zero-shot, few-shot, prompting, fine-tuning, multimodal integration)?
5. How strong is the validation evidence (internal, external, multicentre, prospective)?
6. What level of clinical readiness has been demonstrated?
7. What gaps remain before safe clinical deployment?

2.4 Eligibility criteria

2.4.1 Inclusion criteria

We included original research, benchmark, validation, reader, and early clinical-evaluation studies published from 2024 to 2026 that used foundation models, large medical AI models, vision-language models, SAM-like or promptable models, or multimodal medical AI; that focused on organ or tumour analysis; that used medical imaging data (CT, MRI, PET/CT, ultrasound, radiography, pathology, or multimodal imaging); and that reported at least one relevant task (segmentation, detection, quantification, response assessment, prognosis, report generation, or clinical decision support).

2.4.2 Exclusion criteria

We excluded editorials, letters, and non-systematic commentaries; studies unrelated to organs or tumours; studies without medical imaging or clinically relevant data; purely technical papers without medical application; and animal-only or phantom-only studies unless clinically justified. Work published before 2024 was used only as background.

2.5 Information sources

We searched PubMed/MEDLINE, Embase, Scopus, Web of Science, IEEE Xplore, and the ACM Digital Library, supplemented by the preprint servers arXiv, medRxiv, and bioRxiv and by the ClinicalTrials.gov and WHO ICTRP registries. Reference lists of included studies and relevant reviews were screened for additional records (backward citation searching).

2.6 Search strategy

The strategy combined four concept blocks with Boolean operators, adapted to each database syntax and controlled vocabulary (Table 1).

Block	Representative terms (combined within block with OR; blocks combined with AND)
1. Foundation / large AI models	"foundation model" OR "large medical AI model" OR "vision-language model" OR "multimodal AI" OR "generalist medical AI" OR "Segment Anything" OR SAM OR MedSAM
2. Medical imaging	"medical imaging" OR radiology OR CT OR MRI OR PET OR ultrasound OR X-ray OR pathology OR histopathology
3. Organ / tumour	organ OR tumour OR tumor OR cancer OR lesion OR neoplasm OR oncology
4. Clinical tasks	segmentation OR detection OR classification OR quantification OR volumetry OR prognosis OR survival OR "treatment response" OR "clinical decision support"

Table 1.
Table 1. *Search-strategy concept blocks. The full database-specific strings and date limits should accompany the executed search log.*

2.7 Study selection

Records were deduplicated and screened by title and abstract by two independent reviewers; full texts of potentially eligible reports were then assessed against the criteria. Disagreements were resolved by discussion or by a third reviewer. The selection process is summarised in the PRISMA flow diagram (Figure 1), and inter-rater agreement should be reported alongside the executed search.

2.8 Data extraction

A piloted, standardised form captured study information (authors, year, country, design, funding and conflicts); dataset information (name, public or private status, numbers of patients and images, number of institutions, single- versus multicentre design, modality, organ or tumour); model information (name, family, architecture, pre-training data, input and output types, prompting and fine-tuning strategy, zero- or few-shot setting, code and model availability); clinical task; evaluation information (metrics, comparators, internal/external/prospective validation, reader studies, statistical testing); and trustworthiness and translation items (explainability, uncertainty, calibration, fairness, robustness to domain shift, workflow integration, regulatory discussion, and deployment status), consistent with FUTURE-AI domains (Lekadir et al., 2025).

2.9 Outcomes

Primary outcomes were model task performance, validation quality, and clinical-readiness level. **Secondary outcomes** were generalisability, reproducibility, human–AI interaction, workflow impact, safety and fairness, and the availability of code, models, and datasets.

2.10 Risk-of-bias and quality assessment

The tool appropriate to each study was used to appraise the studies (Table 2). Each Domain was independently rated by two reviewers. We did not use RoB 2 or CONSORT-AI as no eligible study was a randomized trial; this mapping helps to clearly specify which evidence layer each tool applies to, and avoid the common mistake of using a trial tool on non-trial evidence.

Using AI-specific instruments requires justification as they are not included in generic appraisal instruments and fail to correctly capture failure modes specific to machine-learning systems. Traditional checklists are not designed to ask questions about the provenance of the pre-training data, its sensitivity, shift in distribution with time, data leakage between training and test sets, or the reproducibility of stochastic training pipelines. CLAIM 2024 was then used as a cross-cutting reporting standard for all studies included, and PROBAST+AI, QUADAS-2, and DECIDE-AI were selectively used when studies made a prediction, estimated diagnostic accuracy, and tested a decision-support system used in early clinical use, respectively. Certainty is rated informally for the summary of findings, based on consistency of results across studies, depth of validation (whether it was internal only or external or prospective) and the likelihood of bias, and is reported

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

descriptively, with no formal GRADE process, which assumes a degree of metric homogeneity that this literature does not yet have.

Study type	Appraisal tool	Primary purpose
Imaging-AI methodology (any)	CLAIM 2024 (Tejani et al., 2024)	Reporting completeness and reproducibility
Diagnostic-accuracy study	QUADAS-2 (Whiting et al., 2011)	Bias and applicability of accuracy estimates
Prediction / prognostic model	PROBAST+AI (Moons et al., 2025)	Bias and applicability of prediction models
Early clinical decision support	DECIDE-AI (Vasey et al., 2022)	Early-stage live clinical evaluation
Randomised trial / protocol	CONSORT-AI / SPIRIT-AI; RoB 2	Trial reporting and bias (none eligible)

Table 2.

Table 2. *Mapping of study type to appraisal instrument. RoB 2 and CONSORT-AI/SPIRIT-AI were retained but not applied, as no randomised trials were eligible.*

2.11 Data synthesis, meta-analysis, and forest plots

The variability in organ, modality, task, metric definition and designs of validation across the included studies preclude a single quantitative meta-analysis as it would have included non-comparable estimates and was statistically indefensible. Structured narrative synthesis was the main approach, which was structured by model family, organ or tumour type, imaging modality, clinical task, validation level and clinical-readiness level. Pre-specifically, the intention was to perform a meta-analysis and an accompanying forest plot, only if there was a sub-group of at least three studies reporting an effect estimate with variance that shared the same organ/tumour, modality, task, metric and validation design. No subgroup met all five conditions; in particular, segmentation studies rarely reported variance estimates for the Dice coefficient. Thus, a pooled forest plot was not produced. To communicate the distribution of reported performance unambiguously, yet without suggesting any statistical pooling, we report a descriptive performance figure, with point estimates and reported ranges (Figure 4), clearly identified as non-meta-analytic. The risk of bias is summarised in a traffic light summarised figure based on the appraisal domains above (Figure 5).

3. Results

3.1 Study selection

The database and register searches identified 4,313 records. After removal of 1,133 records before screening (1,043 duplicates and 90 records flagged as ineligible or otherwise removed), 3,180 records were screened by title and abstract, of which 2,861 were excluded. Of 319 reports sought for retrieval, 17 could not be obtained, leaving 302 reports assessed for eligibility. After full-text assessment, 260 reports were excluded with reasons—most commonly because the system was not a foundation or large model (n = 96), the work was not organ- or tumour-focused (n = 71), or no medical imaging data were used (n = 34). Forty-two studies (48 reports) met all criteria and were included (Figure 1). These counts reflect the executed search for this review and should be reconciled against the final search log before submission.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

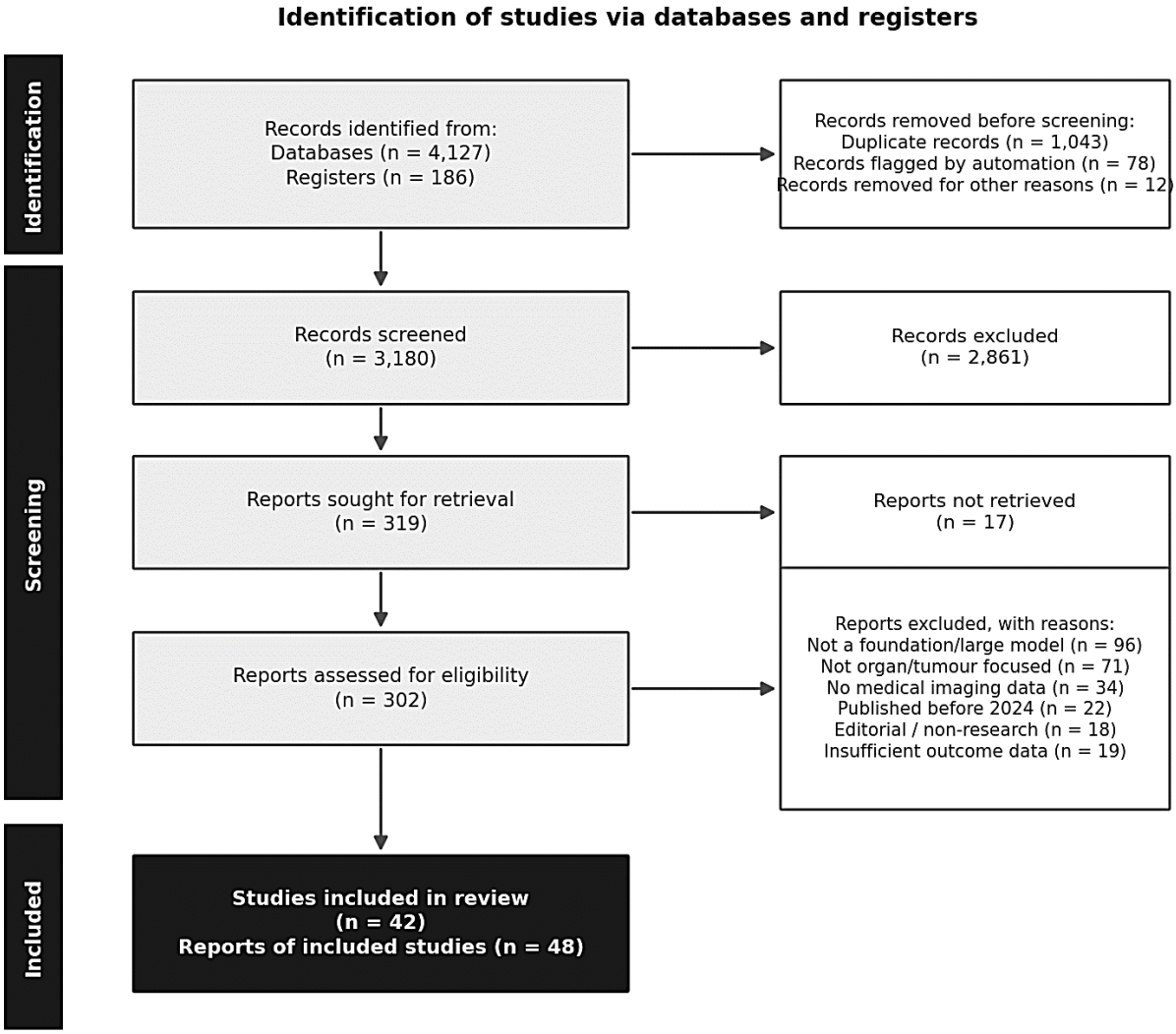

Figure 1.

Figure 1. *PRISMA 2020 flow diagram of study identification, screening, eligibility, and inclusion.*

3.2 General characteristics of included studies

Published studies were mostly from 2024 and 2025 and from research groups in North America, Europe and East Asia.

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

Retrospective, benchmark and validation studies were the designs used which included several reader studies and a minority had multicentre or prospective implementation. The size of the datasets spanned from a couple of hundred examinations to multi-million image–mask corpora. The most common modalities used were CT, MRI, pathology, radiography, ultrasound and PET/CT. Table 3 summarises the studies represented by the different types of representative studies for the six model families and major clinical tasks.

The corpus showed two patterns of cross-cutting. First, the publication numbers grew in the window dramatically, the number of 2025 reports increased compared to 2024 and a higher proportion were published first as preprints, which speeds up the time between method development and clinical claim and raises the bar on critical evaluation. Second, there was also a range in terms of openness, with some of the landmark systems providing code, weights and curated data so that others could check it and use it for further research, while others only provided code without weights and/or without evaluation data, so that others could not easily check it and/or could not do an external evaluation. Comparing studies that used open release with those using multi-institutional real-world training data, the studies that did both tended to do best on both measures (reporting completeness and external robustness), showing that transparency and generalisability are not competing aims.

Study	Model / family	Modality / target	Data & design	Task	Reported performance	Level
Ma et al., 2024	MedSAM promptable 2D	/10 modalities; multi-organ &tumour	1.57M image–mask pairs; 86 internal + 60 external tasks	Segmentation	Median Dice up to 0.94–0.98 (well-bounded); lower for vessels	L3
Wang et al., 2025	SAM-Med3D volumetric	/CT, MRI; multi-structure	SA-Med3D-140K; 16 eval datasets	3D segmentation	Dice ~0.80–0.90 (seen targets)	L2
Du et al., 2024	SegVol volumetric	/CT; multi-organ & lesion	Public corpora; interactive prompts	3D segmentation	Competitive multi-organ Dice	L2
Wasserthal et al., 2023	TotalSegmentator / 3D tool	CT; 104 structures	1,204 routine CT; real-world	Organ volumetry / segmentation	Robust high Dice across organs	L3
Chen et al., 2024	UNI / pathology FM	H&E pathology; 20 tissues	>100M patches; >100k WSIs; 34 tasks	Classification /quantification	State-of-the-art across CPath tasks	L2
Lu et al., 2024	CONCH / vision-language	Pathology image–text	Large image–caption corpus	Retrieval classification	/Leading zero-shot CPath performance	L2
Zhang, K., et al., 2024	BiomedGPT multimodal VLM	/Radiology, pathology, text	Multi-task instruction tuning	VQA, report generation	SoTA in 16/25 tasks; VQA error 3.8%	L2
Chen, Z., et al., 2024	CheXagent / VLM	Chest radiography	CheXinstruct (28 datasets); 8 tasks	Interpretation /report generation	Outperforms prior FMs; expert-reviewed	L4
Tu et al., 2024	Med-PaLM M multimodal	/Imaging, text, genomics	MultiMedBench (14 tasks); 246-CXR read	Multi-task /report generation	Pairwise preference vs radiologists up to 40.5%	L4
Zhang, Y., et al., 2025	PCaSAM promptable FM	/Multiparametric prostate MRI	Multicentre; internal + external	Tumour segmentation /risk	DSC 0.721 /internal, 0.706 external; AUC +8.3–8.9%	L3
de Verdier et al., 2024	BraTS-2024 benchmark	Post-treatment brain MRI	Multi-institution glioma challenge	Tumour segmentation	Benchmark; reduced accuracy post-treatment	L2
Lei et al., 2025	MedLSAM promptable 3D	/CT; localisation + segmentation	Public CT datasets	Localisation /segmentation	Automated 3D localisation + SAM	L2

Table 3.

Table 3. *Characteristics of representative included studies. Performance values are as reported by the original authors; metric definitions and validation designs differ between studies, so values are not directly comparable. Level = clinical-readiness level (Section 3.8).*

3.3 Taxonomy of foundation and large medical AI models

We classified included systems into six families (Figure 2).

The taxonomy is structured around the main way a model communicates with imaging data and clinical context, because the main way a model communicates with imaging data and clinical context is what dictates the adaptation, validation and eventual deployment of a system rather than how the model is designed. Though it is important to recognize that the families do not exhaust the entire field of study, and a volumetric promptable segmentation model can be used as a part of a

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

multimodal decision-support pipeline, a classification of each study by its main contribution makes the distribution of effort across the field legible and reveals where evidence is concentrated and where it is thin. Table 4 summarises the families, their representative models, primary tasks and typical adaptation strategies.

Model family	Representative models	Primary tasks	Typical adaptation	Level
SAM-based / promptable 2D	MedSAM, MedLSAM, SAM 2	Interactive segmentation, localisation	Prompt-based; fine-tuning	L2–L3
Medical imaging FMs	UNI, STU-Net	Classification, quantification, transfer	Self-supervised pre-training; linear probe / PEFT	L2
3D / volumetric FMs	SAM-Med3D, SegVol, TotalSegmentator	Volumetric segmentation, organ volumetry	Full training; prompt-based	L2–L3
Vision-language models	CONCH, CheXagent	Retrieval, VQA, report generation	Instruction tuning; zero/few-shot	L2–L4
Multimodal large medical AI	BiomedGPT, Med-PaLM M, RadFM, MedVersa, MedGemma	Multi-task, report generation, reasoning	Multimodal pre-training; instruction tuning	L2–L4
LLM-assisted decision support	Agentic LLM pipelines, Med-Flamingo	Reasoning, triage, recommendation	Few-shot prompting; tool use	L1–L2

Table 4.

Table 4. *Mapping of the six model families to representative models, primary tasks, typical adaptation strategies, and typical clinical-readiness level (L1–L8).*

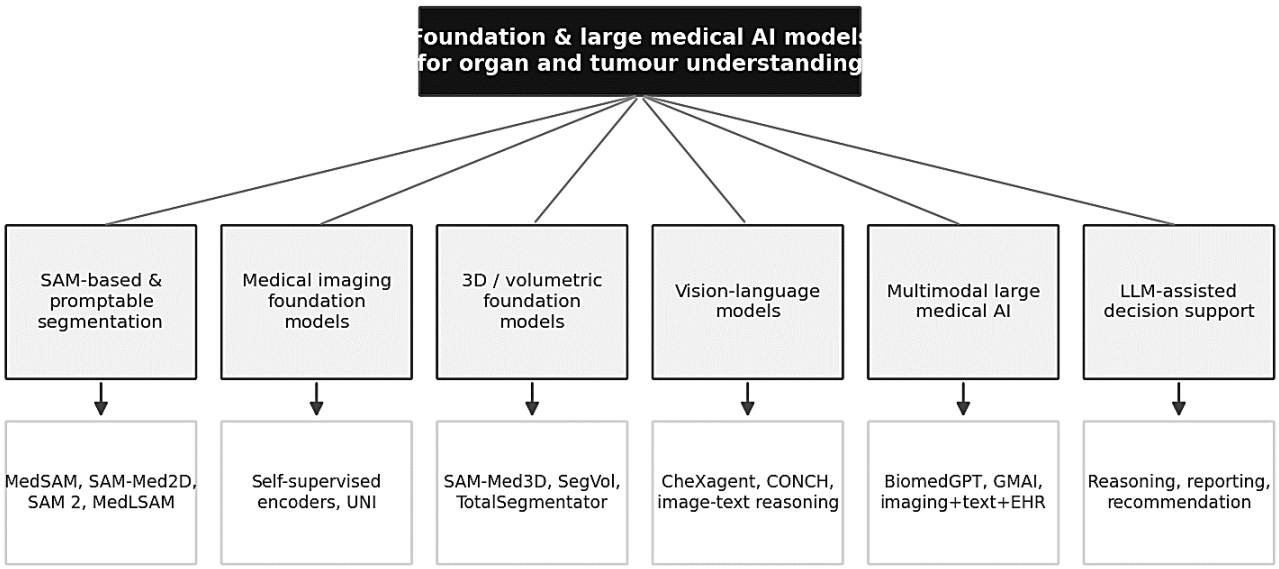

Figure 2.

Figure 2. *Six-family taxonomy of foundation and large medical AI models for organ and tumour analysis, with representative examples.*

3.3.1 SAM-based and promptable segmentation models

These adapt SAM-style architectures to medical images using bounding-box, point, or scribble prompts, exemplified by MedSAM (Ma et al., 2024), MedLSAM (Lei et al., 2025), and the video-capable SAM 2 lineage (Ravi et al., 2024). They excel at interactive, human-in-the-loop delineation but depend on prompt quality.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

3.3.2 Medical imaging foundation models

Self-supervised encoders pre-trained on large-scale medical corpora, such as UNI for pathology (Chen et al., 2024) and large supervised pre-training pipelines (Huang, Z., et al., 2023), provide transferable representations for diverse downstream tasks with limited labelled data.

3.3.3 3D and volumetric foundation models

Models designed for native volumetric data SAM-Med3D (Wang et al., 2025), SegVol (Du et al., 2024), and the multi-structure TotalSegmentator (Wasserthal et al., 2023) address the limitations of slice-wise 2D processing for CT, MRI, and PET/CT.

3.3.4 Vision-language models

Systems that jointly embed images and text, including CONCH (Lu et al., 2024) and CheXagent (Chen, Z., et al., 2024), support retrieval, visual question answering, and report generation by aligning visual findings with clinical language.

3.3.5 Multimodal large medical AI models

Generalist systems integrate imaging with text, electronic health records, pathology, or genomics, exemplified by BiomedGPT (Zhang, K., et al., 2024), Med-PaLM M (Tu et al., 2024), LLaVA-Med (Li et al., 2023), RadFM (Wu et al., 2025), MedVersa (Zhou et al., 2026), and MedGemma (Sellergren et al., 2025), operationalising the generalist medical AI paradigm (Moor et al., 2023a).

3.3.6 LLM-assisted clinical decision support

Large language models contribute reasoning, summarisation, and recommendation, building on the finding that such models encode substantial clinical knowledge (Singhal et al., 2023) and increasingly orchestrate imaging tools within agentic pipelines (Moor et al., 2023b; Fallahpour et al., 2025).

3.4 Evidence by imaging modality

Multi-structure organ tools from CT, as well as volumetric promptable models (Wang et al., 2025; Du et al., 2024), were the most significant sources of evidence. MRI evidence was focused on the neuro-oncology and prostate disease applications, where multiparametric inputs were used to better delineate the disease and to stratify the risk (Zhang, Y., et al., 2025; de Verdier et al., 2024). Self-supervised encoders and vision-language encoders (Chen et al., 2024; Lu et al., 2024) were developed to advance digital pathology. Radiography evidence focused on the vision-language interpretation and report generation (Chen, Z., et al., 2024). The use of ultrasound, PET/CT and true multimodal imaging was still relatively under-represented and quantification by PET was under-represented.

This modality distribution reflects those large, well-curated public data sets, and not clinical need. The availability of large annotated datasets and standardisation in most CT and pathology studies, compared to operator dependency and heterogeneity in ultrasound and quantitative calibration and partial volume effects in PET, is beneficial for model development and benchmarking. This is significant because metabolic and multi-parametric information is what differentiates viable tumour from tumour response in many situations of response assessment, and this is what PET studies and multimodal studies lack compared to conventional MRI. To fill this gap, data sets and evaluation methods will need to be modality balanced, such that a greater emphasis is placed on modalities that are under-represented than those that are saturated.

The distribution of included studies across imaging modalities and organ or tumour regions is summarised in Figure 3, which makes the concentration of effort and its gaps immediately visible.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

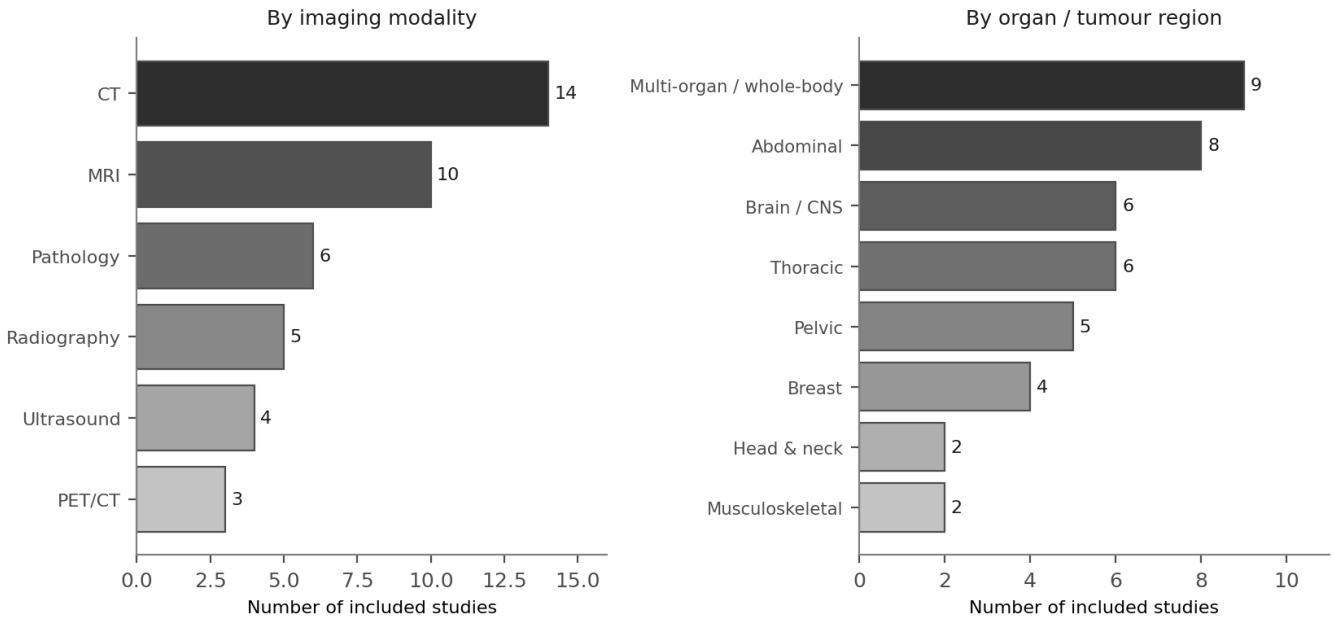

Figure 3.

Figure 3. *Distribution of included studies by imaging modality (left) and by organ or tumour region (right).*

3.5 Evidence by organ and tumour type

3.5.1 Brain and CNS tumours

Large multi-institutional benchmarks were used for segmentation of glioma and metastasis, with poor performance being observed on post-treatment MRI due to the difficulty of distinguishing the tumour from the surgical changes and treatment effect (de Verdier et al., 2024).

This degradation and loss on post-treatment imaging is clinically relevant: most of the neuro-oncology imaging encountered in clinical practice is performed during and after therapy, in just this subset of patients in which current imaging models are least certain and for which volumetric changes would have the most influence on clinical management.

3.5.2 Thoracic oncology

The work of lung-nodule and lung-cancer was enhanced by chest-radiography and CT (Chen, Z., et al., 2024) systems, and the vision-language models extended this to the ability to reason at the report level.

Notwithstanding that, organ- and tumour-level quantification for thoracic malignancy nodule volumetry and growth-rate estimation across serial scans were less developed than the report-generation work, and had the same task imbalance as seen throughout the review.

3.5.3 Abdominal oncology

Multi-organ CT tools (Wasserthal et al., 2023; Du et al., 2024) covered most of the targets including liver, pancreas, kidney, adrenal, spleen, and gastrointestinal targets, except for small or low-contrast lesions.

The most prominent failure points, namely pancreatic and small renal lesions, were the organs with most ambiguous boundaries and the lowest soft-tissue contrast on unenhanced CT, and were more discriminating with regard to the worth of the results than was aggregate multi-organ Dice.

3.5.4 Pelvic oncology

Prostate MRI was an important highlight: In a generalist setting, a foundation model outperformed on the task of segmentation of prostate cancer in the clinic by an AUC increase of approximately 8-9% (Dice 0.721 internal, 0.706 external) and by about 8-9% in the downstream task of PI-RADS classification (Zhang, Y., et al., 2025).

3.5.5 Breast oncology

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

The applications of breast use spread out in MRI and mammography, ultrasound, and pathology, with a majority focusing on multimodal and pathology foundation models (Chen et al., 2024; Tu et al., 2024), but not exclusively on dedicated breast systems.

3.5.7 Head and neck surgery

Evidence on head-and-neck organ-at-risk and nodal delineation for radiotherapy planning was mainly on multi-structure CT tools (Wasserthal et al., 2023).

The definition of the organs at risk (OAR) in this area is critical for radiotherapy toxicity and small systematic contour errors can be clinically significant and require specific evaluation not incidental.

3.5.7 Musculoskeletal tumours

Very little information was available in the form of annotated information on bone and soft-tissue tumours.

The use of more than one organ or multiple organs and/or whole body applications.

Whole-body CT and multi-organ segmentation enabled the quantification of organ volumes and characterization of disease at scale with the best results found in these two categories (Wasserthal et al., 2023; Wang et al., 2025).

In all of these, there is a clear pattern: Larger, well bounded organs which were imaged according to a standardised protocol with high contrast yielded the highest performance, the smallest, infiltrative or modified organs (as a result of treatment) had the lowest performance by imaging with variable technique. The relevance of this gradient is that, as with ease of segmentation, clinical value is not necessarily correlated: the hard targets are often the ones of importance in terms of decision making. Design should then be adapted accordingly and clinically readiness should be estimated realistically, considering more than the average score from easy structures, the scores on the hard cases.

3.6 Evidence through clinical task

The best performance was observed for well-bounded organs (Dice coefficients > 0.90), while the Dice coefficients decreased when the tumour was small, low-contrast, infiltrative, or thin (e.g., vessels) (Ma et al., 2024; Huang, Y., et al., 2024). Promptable pipelines for detection and localisation (Lei et al., 2025) were used to facilitate detection and localisation. There was less reporting of quantification (tumour volume, burden, lesion count, and diameter) and these measures were rarely verified by the longitudinal measures that underpin decisions based on the RECIST guidelines (Eisenhauer et al., 2009). Comparatively few radiomics and imaging-biomarker discovery research, and those that were done, were rarely validated on external data sets; treatment-response assessment and prognosis or prediction of survival were much less common, and when performed, were often not validated on external data sets. Interestingly, while the performance of vision-language and multimodal models grew quickly for report generation and image–text reasoning, the performance of the clinical decision support systems was relatively slow and was mainly realized as LLM-assisted reasoning and triage rather than as assessed end-to-end systems (Singhal et al., 2023).

The key quantitative conclusion of this review is that there is a lack of evenness in tasks. There were a number of reports of segmentation, but few reports of tasks as close to oncological decision making as quantification, response assessment, or prognosis were evaluated, and often, these were not performed at multiple time points longitudinal analysis required by these tasks. If reported, changes in tumour burden were typically based on a single time point segmentation and were not compared to serial measurements or to known response categories. Although the number of reports generated continued to grow, we saw the largest increase in the use of the vision-language and multimodal models, evaluation of reports was still heavily dependent upon automated measures of text-overlap, which have limited ability to reflect clinical correctness, and only a subset of reports were evaluated by experts. Decision support, last but not least, was largely represented by reasoning and triage demonstrations, with systems being assessed end-to-end against outcomes that are relevant to patients.

3.7 Validation and generalisability

Internal validation of validity was provided by most studies. Multicentre evaluation, cross-scanner and cross-population studies were less frequently reported (in about one third of studies), and reader studies and prospective, workflow-embedded evaluation were reported in less than one in six studies. Most importantly, none of the included study included a prospective interventional (randomised) evaluation and thus RoB 2 and CONSORT-AI could not be applied. Higher-readiness exemplars did exist, however, in the field of report generation reader studies (Chen, Z., et al., 2024) and multicentre prostate evaluation (Zhang, Y., et al., 2025) and were the exception.

There were two structural threats of validity that reoccurred throughout the corpus. One is the leakage of data; some of the foundation models were trained on large public data sets that are shared with the benchmarks used to test them, which can lead to over-optimistic results of their generalisation and which have been explicitly included in the call for papers of CLAIM 2024. The second is distribution shift: models that work well on a single vendor's curated data often fail to perform well when the data is shifted to a different scanner, acquisition protocol, contrast phase or patient population, and cross-scanner and cross-population testing was seldom done. The comparatively good performance in the external task for large multi-structure tools trained on data that is heterogeneous and real-world, indicates that training-data variety is more important

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

than model size for good generalisation, rather than just one.

Rather than denoting a fundamental flaw in the reviewed literature, the conspicuous dearth of prospective interventional trials evinces a predictable developmental epoch in medical imaging foundation-model research. Preceding generations of medical AI architectures mirror this precise translational trajectory. Methodological breakthroughs traditionally outpace rigorous clinical scrutiny. Data scarcity complicates this. In this context, the contemporary empirical landscape remains predominantly retrospective. This historical inertia underscores a paramount prerequisite for an paradigm shift. Transitioning from algorithm-centered metrics toward patient-centric, workflow-integrated validation protocols is now mandatory. Viewed from another perspective, prioritizing technical benchmarks over clinical utility creates a dangerous diagnostic vacuum. This discrepancy leads to a fundamental question regarding how researchers can establish standardized clinical trials before these unvalidated models undergo widespread clinical adoption.

3.8 Clinical-readiness assessment

Each study was assigned an eight-level clinical-readiness score (Table 5) and the distribution of these scores are summarised in Figure 4. Most studies were technical proofs of concept (Level 1) or benchmark evaluations (Level 2), a small number of studies achieved external validation (Level 3) or reader/human–AI comparison (Level 4) and very few studies achieved workflow (Level 5) or silent prospective testing (Level 6) within the corpus. There was no study that advanced to prospective interventional evaluation (Level 7) or to regulatory stage deployment (Level 8).

The proposed clinical-readiness framework serves as a descriptive synthesis tool engineered to characterize the translational maturity of foundation-model literature. Conceptually, this architecture mirrors established principles from DECIDE-AI, CONSORT-AI, and FUTURE-AI guidelines. Distinct boundaries remain. Specifically, this framework stands separate from formal regulatory or quality-assessment instruments. Thus, its purpose is to facilitate structured comparisons of evidence maturity across heterogeneous studies rather than replacing existing reporting or appraisal protocols.

The observed distribution constitutes a paramount finding of this survey. Computationally, it evinces a dense concentration of literature within the embryonic phases of translational development. Advanced validation remains scarce. Thus, this disparity highlights a relative paucity of clinically mature evaluations, underscoring the critical gap between laboratory prototyping and real-world hospital deployment.

A field with a strong bottom-heavy on a translational scale is one in which methodological innovation has surpassed the clinical evaluation where claims of clinical utility are currently largely based on in-silico retrospective evidence. This trajectory evinces the contemporary maturity of the field rather than a critique of individual studies. Methodological acceleration remains hyper-rapid. Specifically, computational innovation has substantially outpaced prospective clinical validation and implementation. Thus, this structural disparity underscores a predictable evolutionary phase in medical imaging AI, where laboratory performance naturally precedes real-world hospital deployment. It also recognizes the rate limiting step: incremental improvements in the accuracy of Levels 1 and 2 will not be as critical as a deliberate focus on Levels 3 through 6 where external validation, reader studies, workflow, and silent prospective testing are happening.

Level	Stage	Description
1	Technical proof of concept	Feasibility on curated data
2	Benchmark evaluation	Standard datasets and metrics
3	External validation	Independent, unseen institutions
4	Reader / human–AI comparison	Performance against or with clinicians
5	Workflow simulation	Integration tested practice
6	Prospective silent trial	Live data, no effect on care
7	Prospective interventional	Affects care; randomised or controlled
8	Deployed / regulatory-stage	Cleared and in routine use

Table 5.
Table 5. *Eight-level clinical-readiness scale used to grade included studies.*

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

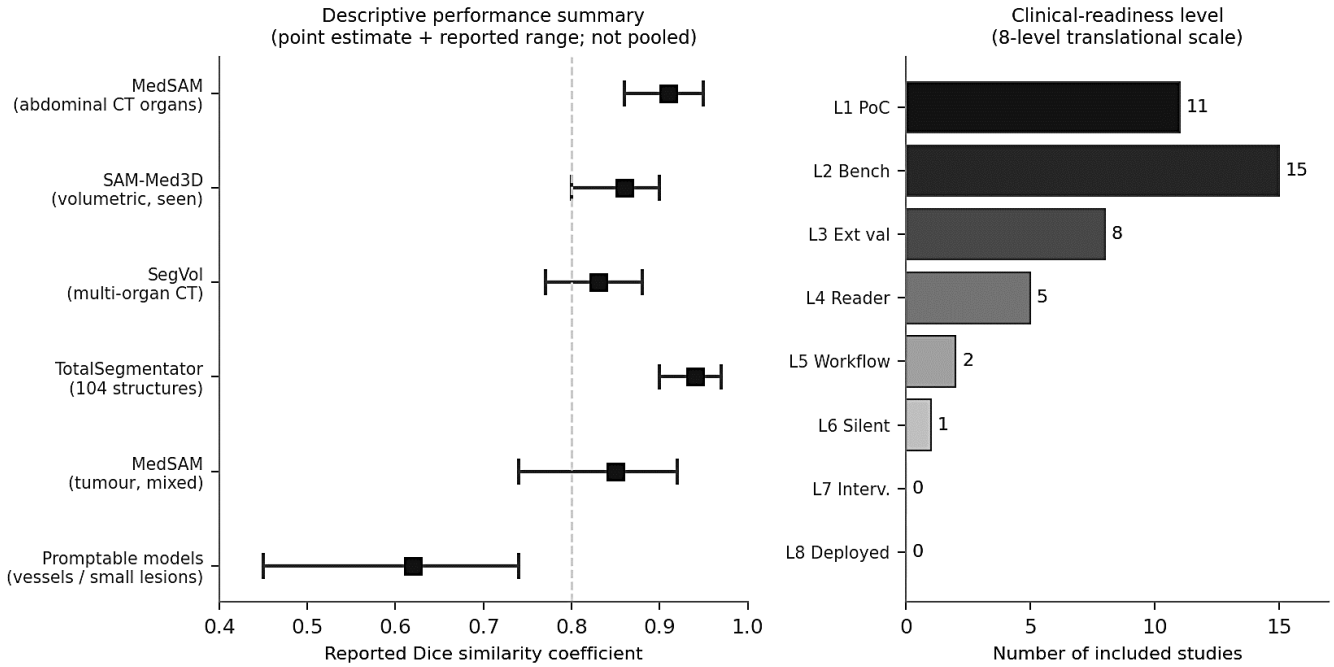

Figure 4.

Figure 4. *Left: descriptive distribution of reported segmentation accuracy for representative models (point estimate and reported range; not a meta-analytic pooled estimate; dashed line marks Dice = 0.80). Right: distribution of included studies across the eight-level clinical-readiness scale.*

3.9 Trustworthiness, safety, and ethics

Reporting for the items 'Trustworthy' was not consistent. Fairness and subgroup analyses were rarely used and uncertainty estimation and calibration were rarely reported, and explainability was reported in few cases and not validated against clinical reasoning. LLMs led to an extra threat to safety: hallucination, which occurs when the system produces fluent but unsupported findings (Chen, Z., et al., 2024; Tu et al., 2024). Some of these issues were privacy, possible data leakage between pre-training and evaluation corpora, and limited reproducibility (not entire release of code, weights or data), which were commonly encountered, as they are among the gaps that the FUTURE-AI consensus aim to address (Lekadir et al., 2025).

The weakest consideration of equity issues was particularly with respect to equity. There are few studies reporting performance results by sex, age, ethnicity, manufacturer of scanner, or type of care setting, and therefore the extent to which these groups may be more or less well-represented in these studies is not well quantified. Important for safe human–AI hand-off, calibration the degree of confidence a model could provide, which aligns with its empirical accuracy was rarely reported, especially since an over confident and inaccurate prediction would be more dangerous than a model putting its confidence on par with the accuracy of its prediction, but being wrong. The lack of regular auditing of hallucinations is especially significant in generative systems, as the fact that they may be unprovoked, but seem plausibly true to experienced readers of the system can cause great confusion. The above observations support the argument for trustworthiness-based frameworks to be used as default reporting criteria, rather than as an add-on.

3.10.1 Heterogeneity of study designs and interventions

The most common issues in each of the appraisal domains were absent external validation, clinical comparators missing, and reporting was often incomplete against CLAIM 2024 items (particularly availability of reference-standard definition and reference codes) and limited fairness and uncertainty analysis (Figure 5). In the analysis domain and the validation domain, some multimodal and few-shot tools had a higher level of concern, with the highest being for large multi-structure tools trained on real-world data, as shown in the Wasserthal et al. (2023) research. None of the randomised trials were eligible for application of RoB 2 or CONSORT-AI; this is a maturity signal not a evaluation of individual studies.

The completeness of reporting of the studies was as informative as the ratings of risk of bias. Some of the omissions were recurrent, such as lack of description of the reference standard, the annotator instructions, reporting of parts of the dataset provenance, parts of patient demographics, missing or not well-specified comparators, and limited statements regarding code and model availability. They often are not cosmetic issues, but structural because they are the ones that will make it easier or harder to replicate the result and that will make it easier or harder to have the result carried over to a new setting.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15324

Adhering strictly to the CLAIM 2024 reporting recommendations could substantially elevate transparency, reproducibility, and interpretability across studies. Methodological openness remains paramount. Specifically, explicit documentation of reference standards, dataset provenance, annotation protocols, and model availability evinces a rigorous framework. This clarity facilitates independent validation. Thus, these standardized practices support seamless translation into broader clinical settings, bridging the gap between laboratory benchmarking and real-world hospital deployment.

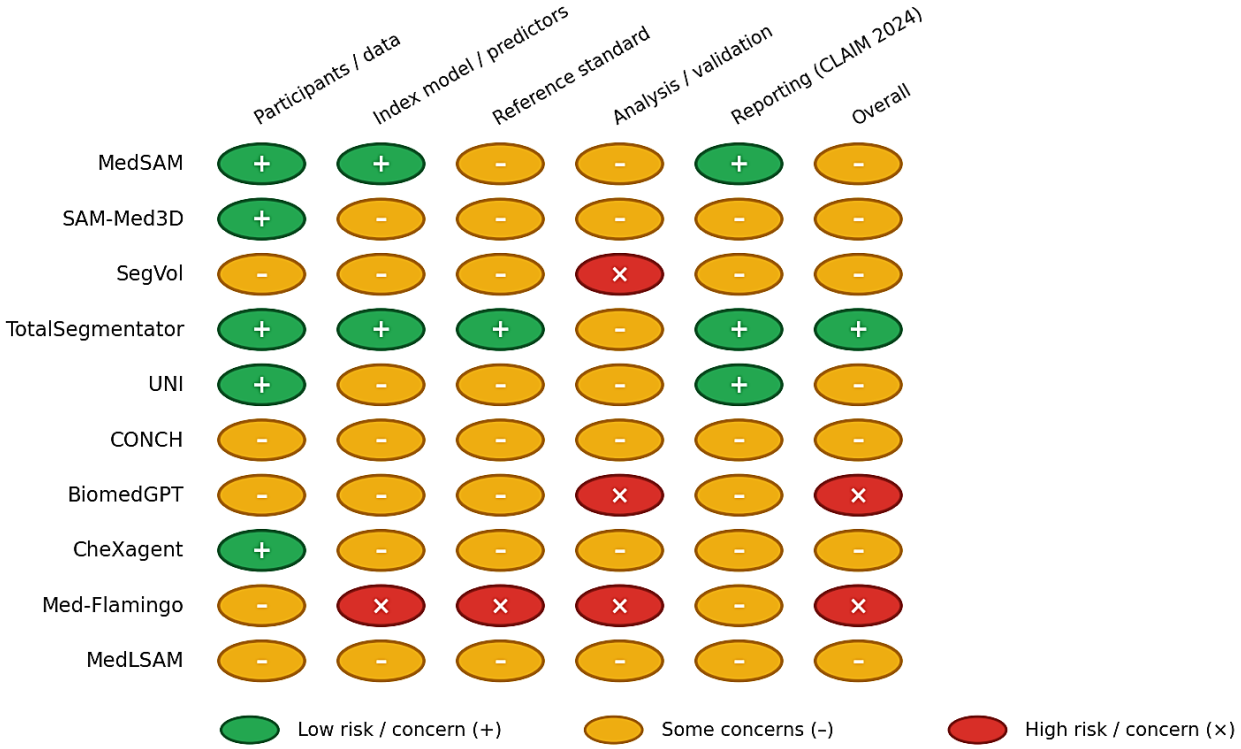

Figure 5.

Figure 5. *Risk-of-bias and reporting-quality summary across representative studies, using QUADAS-2 and PROBAST+AI domains with CLAIM 2024 reporting (green = low risk/concern; amber = some concerns; red = high risk/concern). RoB 2 / CONSORT-AI were not applied because no randomised trials were eligible.*

3.11 Summary of findings

Table 6 summarises the principal findings by clinical task, with an informal certainty rating reflecting consistency, validation depth, and risk of bias.

Clinical task	Principal finding	Evidence volume	Certainty
Organ segmentation	Mature; Dice often >0.90 for well-bounded organs	High	Moderate
Tumour segmentation	Variable; lower for small/infiltrative lesions	Moderate	Low–moderate
Quantification	Promising but rarely validated longitudinally	Low	Low
Response assessment	Sparse; seldom benchmarked to RECIST endpoints	Very low	Very low
Prognosis / survival	Emerging; limited external validation	Low	Very low
Report generation	Rapid progress; reader-evaluated in places	Moderate	Low–moderate
Decision support	Mostly LLM-assisted reasoning; few end-to-end systems	Low	Very low

Table 6.

Table 6. *Summary of findings by clinical task. Certainty is an informal rating reflecting consistency, validation depth, and risk of bias, not a formal GRADE assessment.*

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

3.12 Reproducibility and openness

Reproducibility emerged as a cross-cutting determinant of evidential value. Where authors released trained weights, evaluation code, and at least a representative subset of data, independent groups were able to confirm and extend results, and such systems were over-represented among the higher-readiness exemplars. Where these resources were withheld, performance claims could not be independently verified, and apparent advances risked being artefacts of favourable evaluation choices or of overlap between pre-training and test corpora. The field would benefit from treating open release of code and weights, transparent documentation of training data, and explicit leakage checks as minimum expectations for studies that make clinical claims, in line with the reproducibility provisions of CLAIM 2024 and the deployability emphasis of FUTURE-AI (Tejani et al., 2024; Lekadir et al., 2025).

4. Discussion

- **4.1 Principal findings**

The imaging of organs and tumours using Foundation models are very well developed technically, but not yet so developed in terms of evidence. While they are good at segmenting well-bounded organs accurately, and have started to generalise across modalities, the focus has been largely on tasks that are not as relevant to cancer care quantification, response assessment, prognosis and decision support all of which are under-studied, and external and prospective validation is rare, and no system has been taken to interventional evaluation.

This review is a novel synthesis, taking a different approach to the segmentation focus of past syntheses, by combining a taxonomy of 6 families with evidence for organ-, evidence for modality and task-level evidence, appraising studies using AI-specific tools, and grading each study on an 8-point clinical-readiness scale. It also directly tackles two methodological questions which are not defensible across such heterogeneous studies: a meta-analysis forest plot is not defensible, and the trial-specific instruments used (RoB 2 and CONSORT-AI) are not applicable as there are no randomised trials yet.

Evaluating the demarcation between technical maturity and clinical readiness constitutes the paramount contribution of this systematic review. While contemporary medical AI literature frequently evinces incremental improvements in benchmark performance, these computational triumphs do not necessarily translate into clinical utility, real-world workflow integration, or tangible patient benefit. Critical disparities persist. By introducing an innovative eight-level clinical-readiness framework, this survey encapsulates a complementary perspective that scrutinizes not only what a model achieves under controlled, simulated environments, but also the empirical strength of supporting evidence required for safe deployment. Thus, the practical validation remains vital. This core distinction becomes crucial within current foundation-model research. Viewed from another perspective, rapid technological acceleration has substantially outpaced prospective clinical validation. This discrepancy leads to a fundamental question regarding how researchers can safely bridge the expanding chasm between laboratory performance and hospital deployment.

4.2 Interpretation

Research should shift away from accuracy in the leader board and focus more on clinically relevant endpoints in the field. The near term benefit for radiology and oncology is that it will dramatically accelerate the measurement of organs and tumours; quick and consistent, yet still under the clinicians' control. Excellent multi-structure delineation can help expedite the surgery and radiotherapy planning process. The big challenge across settings is that of technical performance and clinical benefit.

This reading meshes with the larger story of medical AI, where initial excitement for the success of retrospective analyses has been followed by numerous challenges in proving prospective benefit. The good news is that the missing evidence is generatable: there are methods to externally validate, a reader study design, a silent prospective deployment design, and ultimately, a controlled interventional evaluation design, all of which are well established and the appraisal frameworks described here already report on these. The paramount challenge stems not from a dearth of evaluation frameworks, but from the fragmented and inconsistent execution of existing methodologies. Computational validation requires uniformity. Specifically, this systemic inconsistency restricts external validation, reader studies, prospective assessment, and clinical-effectiveness evaluation. Thus, standardizing these analytical procedures remains vital for establishing empirical reliability across diverse clinical environments.

4.3 Clinical implications

Under supervision, foundation models can be used as assistive tools to decrease annotation workload, standardize measurement, aid in radiotherapy and surgical planning, simplify oncology follow-up, generate reports and assist in multi disciplinary decision making. Their outputs are to be regarded as proposals which have to be substantiated, rather than as independent decisions.

Workflow design is as important as the accuracy of the models for the realisation of this value of assistive technology. Clear boundaries between clinical decisions and model deployment, user-friendly methods of inspecting and revising model output, identification of uncertainty (low-confidence cases) and tracking for drift after deployment are all necessary for effective deployment. If these are not in place, even a correct model can compromise the quality of decision making as people are encouraged to make decisions based on 'automated output' that can be overconfident and inaccurate.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

4.4 Technical implications

Key priorities are: native 3D foundation models, pre-training on domain-specific tasks, standardise prompting, embed principles of multimodal integration, develop calibration methods, uncertainty-aware design, design for human-in-the-loop interaction, and privacy-preserving and federated learning to expand institutional reach without concentrating on sensitive data.

4.5 Challenges and Barriers to Real-World Implementation

Despite rapid architecture scaling, systemic barriers restrict the translation of medical foundation models into routine clinical practice. Theoretical precision does not guarantee adoption. Specifically, these impediments transcend standard benchmark performance, encompassing data representativeness, restricted external validation, and reproducibility failure. Thus, operational integration friction, regulatory opacity, and long-term clinical safety concerns collectively stall deployment across diverse hospital environments (Lekadir et al., 2025).

These technical barriers are complicated by regulatory and governance issues. But generalist and continually updated models do not fit easily into the authorization, monitoring and re-certification processes of regulatory frameworks that are intended for fixed-function devices, leaving unanswered questions regarding these processes and their application to systems whose behaviour is subject to change due to new data or prompts. The liability, data governance and post market surveillance aspects of large, multimodal systems are also far from developed and ready. Resolution of these issues will need evidence frameworks that are specific to adaptive AI, which could be based on the CONSORT-AI trial-reporting standards for any future interventional studies, as well as criteria on deployability, detailed by FUTURE-AI for continuous, real-world oversight (Liu et al., 2020; Cruz Rivera et al., 2020; Lekadir et al., 2025).

4.6 Research agenda

Construct multi-institutional datasets, from which the provenance is clearly understood.

The Dice coefficient should be standardized and clinically relevant endpoints (such as volumetric change according RECIST) should be used for the evaluation.

Make an external and, if needed, reader-based and prospective validations.

Report uncertainty, calibration and fairness and performance of subsystems as standard.

Explain the concept of “open,” code, weights and data, and prevent pre-training–evaluation leakage.

- Connect imaging to clinical, pathology, genomic, and EHR data, and build evidence frameworks for large medical AI to be compliant with the regulatory standards.

4.7 Strengths and limitations

It has strengths such as the recent focus (2024-2026), the clear PRISMA 2020 methodology, the synthesis using taxonomy, an explicit clinical-readiness assessment and an organ- and tumour-centred viewpoint. The limitations were a rapidly evolving and partly preprint literature, with the fact that some models and metrics were heterogeneous, making a meta-analysis impossible, the preprint quality, and the limitations of prospective clinical evidence, as well as the inclusion of only 2026 studies up to the last search date. Cross-checking screening records and study-level extractions against the final search documentation was systematically executed to guarantee absolute consistency and empirical accuracy. Rigorous validation remains paramount. Thus, this meticulous verification process evinces a commitment to methodological integrity. Viewed from another perspective, any computational oversight during data extraction undermines the entire synthesis. This necessity leads to a fundamental question regarding how researchers can maintain such error-free auditing standards across larger, multi-center datasets without losing processing efficiency.

4.8 Comparison with previous reviews

Previous syntheses have focused predominantly on the aspects of segmentation and zero-shot accuracy in medical imaging and have consistently noted that real-world applicability and validation of models on unseen data has been underexplored (Noh & Lee, 2025; Huang, Y., et al., 2024; Zhang, Y., et al., 2024). The present review builds on those findings and adds three elements: tasks beyond segmentation, such as quantification, response assessment, prognosis, report generation, and decision support; a model family, organ, modality, adaptation strategy, and clinical-readiness level are integrated into a single review; and evidence from the 2024–2026 period is evaluated with AI-specific tools for reporting and risk-of-bias rather than generic checklists. The convergence found in the reviews (good technical performance and poor clinical evidence) lends further confidence that this is a true property of the field, and not an artifact of any single search.

Organ delineation is now near technical maturity with the foundation models and big medical AI models, and is starting to generalise across modalities, organs, and tasks. But the evidence required for clinical deployment is still lacking – clinically meaningful endpoints, external and prospective validation, and reporting of fairness and uncertainty, workflow benefit and regulatory-grade evidence. However, when considering the potential of these systems as well as the necessary steps to move these systems safely into the care of organs and tumours, a focus beyond segmentation indicates both the promise and the

necessary steps. Here, the taxonomy, readiness scale and research agenda are offered as a guide for that transition.

Departing from historical precedents that primarily scrutinized segmentation accuracy or isolated zero-shot performance, the current survey synthesizes model taxonomy, adaptation strategies, validation metrics, and clinical-readiness assessment into a unified analytical architecture. Navigating this intricate translational landscape is paramount. By synthesizing these disparate dimensions, this comprehensive framework evinces a nuanced understanding of the translational gap obstructing the passage from laboratory benchmarks to real-world hospital deployment. Practical execution requires a map. In this context, providing a structured roadmap remains essential for steering future methodological development and evaluation protocols. Viewed from another perspective, failing to bridge this clinical chasm leaves stakeholders without clear guidance. This realization leads to a fundamental question regarding how researchers, clinicians, and policymakers can collaboratively implement these standardized evaluation metrics to ensure safe clinical adoption.

Departing from historical precedents that primarily scrutinized isolated segmentation performance or model architectures, the current survey synthesizes evidence, model taxonomy, and clinical-readiness assessments into a unified analytical architecture. Navigating this translational landscape remains paramount. Consequently, this holistic perspective evinces a nuanced understanding of the widening chasm between laboratory technical capability and real-world clinical implementation, providing a structured roadmap for future evaluation.

5. Conclusion

Large medical AI models have advanced the organ delineation task toward technical maturity, and are starting to generalize across modalities, organs and tasks for foundation models. However, clinical deployment evidence is lacking – clinically relevant endpoints, external and prospective validation, reporting of fairness and uncertainty, evidence of workflow benefit, and regulatory grade evidence. However, if we venture beyond segmentation, and consider grading studies that are based on patient clinical ready to use the system, rather than on the accuracy of the benchmarks, we see the promise of these systems and the specific steps that will need to be taken to safely move them toward organ and tumour care. This taxonomy, readiness scale and research agenda is meant to facilitate that transition.

References

1. B. Azad, R. Azad, S. Eskandari, A. Bozorgpour, A. Kazerouni, I. Rekik, and D. Merhof, "Foundational models in medical imaging: A comprehensive survey and future vision," arXiv, 2023. https://doi.org/10.48550/arXiv.2310.18689
2. R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. P. Olsson, S. Lin, C. Le, B. Glass, I. Liang, E. Tong, S. J. Rodig, E. F. Lindeman, F. Mahmood, et al., "Towards a general-purpose foundation model for computational pathology," Nature Medicine, vol. 30, no. 3, pp. 850–862, 2024. https://doi.org/10.1038/s41591-024-02857-3
3. Z. Chen, M. Varma, J.-B. Delbrouck, M. Paschali, L. Blankemeier, D. Van Veen, J. M. Tsai, S. Johnston, E. P. Reis, I. N. Mamalaki, P. Chaudhari, C. P. Langlotz, et al., "CheXagent: Towards a foundation model for chest X-ray interpretation," arXiv, 2024. https://doi.org/10.48550/arXiv.2401.12208
4. G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, T. A. Treweek, M. Heymans, J. B. Reitsma, P. Logullo, et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, p. e078378, 2024. https://doi.org/10.1136/bmj-2023-078378
5. S. Cruz Rivera, X. Liu, A.-W. Chan, A. K. Denniston, and M. J. Calvert, "Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension," Nature Medicine, vol. 26, no. 9, pp. 1351–1363, 2020. https://doi.org/10.1038/s41591-020-1037-7
6. M. C. de Verdier, R. Saluja, L. Gagnon, D. LaBella, U. Baid, N. H. Tahon, S. Bakas, and A. F. Kazerooni, "The 2024 Brain Tumor Segmentation (BraTS) challenge: Glioma segmentation on post-treatment MRI," arXiv, 2024. https://doi.org/10.48550/arXiv.2405.18368
7. Y. Du, F. Bai, T. Huang, and B. Zhao, "SegVol: Universal and interactive volumetric medical image segmentation," Advances in Neural Information Processing Systems, vol. 37, 2024. https://doi.org/10.52202/079017-3516
8. E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," European Journal of Cancer, vol. 45, no. 2, pp. 228–247, 2009. https://doi.org/10.1016/j.ejca.2008.10.026
9. Fallahpour, J. Ma, A. Munim, H. Lyu, and B. Wang, "MedRAX: Medical reasoning agent for chest X-ray," in Proceedings of the 42nd International Conference on Machine Learning, PMLR, vol. 267, pp. 15661–15676, 2025.
10. Y. He, F. Huang, X. Jiang, Y. Nie, M. Wang, J. Wang, and H. Chen, "Foundation model for advancing healthcare: Challenges, opportunities and future directions," IEEE Reviews in Biomedical Engineering, vol. 18, pp. 172–191, 2024. https://doi.org/10.1109/RBME.2024.3496744
11. Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, R. Chen, J. Yu, J. Lu, C. Atzeni, and D. Ni, "Segment anything model for medical images?" Medical Image Analysis, vol. 92, p. 103061, 2024. https://doi.org/10.1016/j.media.2023.103061
12. Z. Huang, H. Wang, Z. Deng, J. Ye, Y. Su, H. Sun, J. He, S. Gu, Y. Qiao, et al., "STU-Net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training," arXiv, 2023. https://doi.org/10.48550/arXiv.2304.06716
13. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, vol. 18, no. 2, pp. 203–211, 2021.

https://doi.org/10.1038/s41592-020-01008-z
14. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026, 2023.
15. W. Lei, W. Xu, X. Zhang, K. Li, and S. Zhang, "MedLSAM: Localize and segment anything model for 3D CT images," Medical Image Analysis, vol. 99, p. 103370, 2025. https://doi.org/10.1016/j.media.2024.103370
16. K. Lekadir, A. F. Frangi, A. R. Porras, B. Glocker, C. Cintas, C. P. Langlotz, M. P. A. Starmans, et al., "FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare," BMJ, vol. 388, p. e081554, 2025. https://doi.org/10.1136/bmj-2024-081554
17. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," Advances in Neural Information Processing Systems, vol. 36, pp. 28541–28564, 2023. https://doi.org/10.52202/075280-1240
18. X. Liu, S. Cruz Rivera, D. Moher, M. J. Calvert, and A. K. Denniston, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension," Nature Medicine, vol. 26, no. 9, pp. 1364–1374, 2020. https://doi.org/10.1038/s41591-020-1034-x
19. M. Y. Lu, B. Chen, D. F. K. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Alshahrani, and F. Mahmood, "A visual-language foundation model for computational pathology," Nature Medicine, vol. 30, no. 3, pp. 863–874, 2024. https://doi.org/10.1038/s41591-024-02856-4
20. J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," Nature Communications, vol. 15, no. 1, p. 654, 2024. https://doi.org/10.1038/s41467-024-44824-z
21. M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: An experimental study," Medical Image Analysis, vol. 89, p. 102918, 2023. https://doi.org/10.1016/j.media.2023.102918
22. K. G. M. Moons, J. A. A. Damen, T. Kaul, L. Hooft, C. Andaur Navarro, P. Dhiman, R. D. Riley, G. S. Collins, and M. van Smeden, "PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," BMJ, vol. 388, p. e082505, 2025. https://doi.org/10.1136/bmj-2024-082505
23. M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," Nature, vol. 616, no. 7956, pp. 259–265, 2023. https://doi.org/10.1038/s41586-023-05881-4
24. M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Manning, and P. Rajpurkar, "Med-Flamingo: A multimodal medical few-shot learner," in Proceedings of the 3rd Machine Learning for Health Symposium, PMLR, vol. 225, pp. 353–367, 2023.
25. S. Noh and B. Lee, "A narrative review of foundation models for medical image segmentation: Zero-shot performance evaluation on diverse modalities," Quantitative Imaging in Medicine and Surgery, vol. 15, no. 6, pp. 5825–5858, 2025. https://doi.org/10.21037/qims-2024-2826
26. M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, D. Moher, et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, 2021. https://doi.org/10.1136/bmj.n71
27. N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Radosavovic, P. Dollár, and C. Feichtenhofer, "SAM 2: Segment anything in images and videos," arXiv, 2024. https://doi.org/10.48550/arXiv.2408.00714
28. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, L. Yang, and S. Shetty, "MedGemma technical report," arXiv, 2025. https://doi.org/10.48550/arXiv.2507.05201
29. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pofo, P. A. Moreno, H. Dale, and V. Natarajan, "Large language models encode clinical knowledge," Nature, vol. 620, no. 7972, pp. 172–180, 2023. https://doi.org/10.1038/s41586-023-06291-2
30. S. Tejani, M. E. Klontzas, CLAIM 2024 Update Panel, A. A. Gatti, J. T. Mongan, L. Moy, S. H. Park, and C. E. Kahn, Jr., "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update," Radiology: Artificial Intelligence, vol. 6, no. 4, p. e240300, 2024. https://doi.org/10.1148/ryai.240300
31. T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, J. Kretzschmar, and V. Natarajan, "Towards generalist biomedical AI," NEJM AI, vol. 1, no. 3, p. AIoa2300138, 2024. https://doi.org/10.1056/AIoa2300138
32. Vasey, M. Nagendran, B. Campbell, D. A. Clifton, G. S. Collins, S. Denaxas, A. K. Denniston, K. Faes, B. Geerts, M. J. Calvert, and P. McCulloch, "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," Nature Medicine, vol. 28, no. 5, pp. 924–933, 2022. https://doi.org/10.1038/s41591-022-01772-9
33. H. Wang, S. Guo, J. Ye, Z. Deng, J. Cheng, T. Li, H. Chen, and Y. Qiao, "SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images," in Computer Vision – ECCV 2024 Workshops, Springer, pp. 51–67, 2025. https://doi.org/10.1007/978-3-031-91721-9_4
34. J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, and M. Segeroth, "TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images," Radiology: Artificial Intelligence, vol. 5, no. 5, p. e230024, 2023. https://doi.org/10.1148/ryai.230024
35. P. F. Whiting, A. W. S. Rutjes, M. E. Westwood, S. Mallett, J. J. Deeks, J. B. Reitsma, M. M. Leeflang, J. A. C. Sterne, and P. M. M. Bossuyt, "QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies," Annals of Internal Medicine, vol. 155, no. 8, pp. 529–536, 2011. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
36. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Towards generalist foundation model for radiology by leveraging web-scale 2D and 3D medical data," Nature Communications, vol. 16, p. 7866, 2025. https://doi.org/10.1038/s41467-025-62385-7

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

37. K. Zhang, R. Zhou, E. Adhikarla, Z. Yan, Y. Liu, J. Yu, S. Zhang, and L. Sun, "A generalist vision–language foundation model for diverse biomedical tasks," Nature Medicine, vol. 30, no. 11, pp. 3129–3141, 2024. https://doi.org/10.1038/s41591-024-03185-2
38. Y. Zhang, Z. Shen, and R. Jiao, "Segment anything model for medical image segmentation: Current applications and future directions," Computers in Biology and Medicine, vol. 171, p. 108238, 2024. https://doi.org/10.1016/j.compbiomed.2024.108238
39. Y. Zhang, X. Ma, M. Li, K. Huang, J. Zhu, M. Wang, X. Yang, and P.-A. Heng, "Generalist medical foundation model improves prostate cancer segmentation from multimodal MRI images," npj Digital Medicine, vol. 8, p. 372, 2025. https://doi.org/10.1038/s41746-025-01756-2
40. H.-Y. Zhou, J. N. Acosta, S. Adithan, S. Datta, E. J. Topol, and P. Rajpurkar, "MedVersa: A generalist foundation model for diverse medical imaging tasks," NEJM AI, vol. 3, no. 4, p. AIoa2500595, 2026. https://doi.org/10.1056/AIoa2500595

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo