

Foundation Models for Multimodal Medical Imaging Triage: A Simulated Cross-Site Generalization and Calibration Study

Md Iqbal¹, Dr. Arvind Krishnan²

¹Student, Department of Biomedical Engineering, KBNCE, India.

²Assistant Professor, Department of Computer Science and Engineering, KBNCE, India.

ABSTRACT

This article presents a clearly labeled simulated empirical study spanning Computer Science and Engineering / Biomedical Engineering. A simulated multi-site imaging benchmark represented chest radiography, CT-derived summary features, and short clinical text from four hypothetical hospitals. A task-specific CNN, image foundation model, and multimodal vision-language foundation model were compared under internal validation and cross-site shift. Calibration, macro-F1, abstention accuracy, inference time, and subgroup stability were modeled. The multimodal foundation-model configuration achieved the strongest simulated cross-site macro-F1 and the lowest calibration error, but its computational burden was substantially higher than the task-specific baseline. Confidence-based abstention improved safety under the most severe site shift by routing uncertain cases for specialist review. All numerical observations and empirical outcomes are synthetic and are used for methodological illustration only. No real laboratory, bridge, enterprise, medical institution, patient cohort, or operating infrastructure was studied. The article combines current literature with a distinct analytical structure and topic-specific evidence to demonstrate how a future real-world investigation could be designed and reported.

Keywords: Foundation Models, Medical Imaging, Multimodal AI, Cross-Site Generalization, Calibration, Biomedical Engineering

I. Translational Question

This study addresses a current research problem in Computer Science and Engineering / Biomedical Engineering. The central objective is to evaluate performance while retaining attention to durability, safety, reliability, calibration, cost, and deployment constraints.

Because no real dataset was supplied, all quantitative analysis is simulated. The simulation is not evidence that the study was conducted; it is a transparent demonstration of a plausible empirical design.

II. Multimodal Data Configuration

A simulated multi-site imaging benchmark represented chest radiography, CT-derived summary features, and short clinical text from four hypothetical hospitals. A task-specific CNN, image foundation model, and multimodal vision-language foundation model were compared under internal validation and cross-site shift. Calibration, macro-F1, abstention accuracy, inference time, and subgroup stability were modeled.

Variables were selected to represent the principal physical, computational, or biomedical mechanisms relevant to the problem. Bounded variability and controlled noise were introduced to avoid unrealistically deterministic results.

III. Foundation-Model Adaptation

The analytical protocol separated parameter selection from final evaluation where relevant. Inputs were screened for impossible combinations and unit errors. The evaluation included a baseline, an intervention or advanced configuration, and at least one degraded or shifted condition.

Translation from simulation to practice requires clearly defined acceptance criteria, calibration procedures, maintenance ownership, and safe fallback behavior. A pilot should verify performance under realistic disturbances rather than only nominal conditions.

IV. Cross-Site Validation

The multimodal foundation-model configuration achieved the strongest simulated cross-site macro-F1 and the lowest calibration error, but its computational burden was substantially higher than the task-specific baseline. Confidence-based abstention improved safety under the most severe site shift by routing uncertain cases for specialist review.

Table 1. Simulated primary comparative outcomes.

Model	Internal F1	Cross-site F1	Calibration error	Compute index
Task-specific CNN	0.91	0.78	0.082	100
Image foundation model	0.93	0.86	0.061	168
Multimodal FM	0.94	0.89	0.044	221

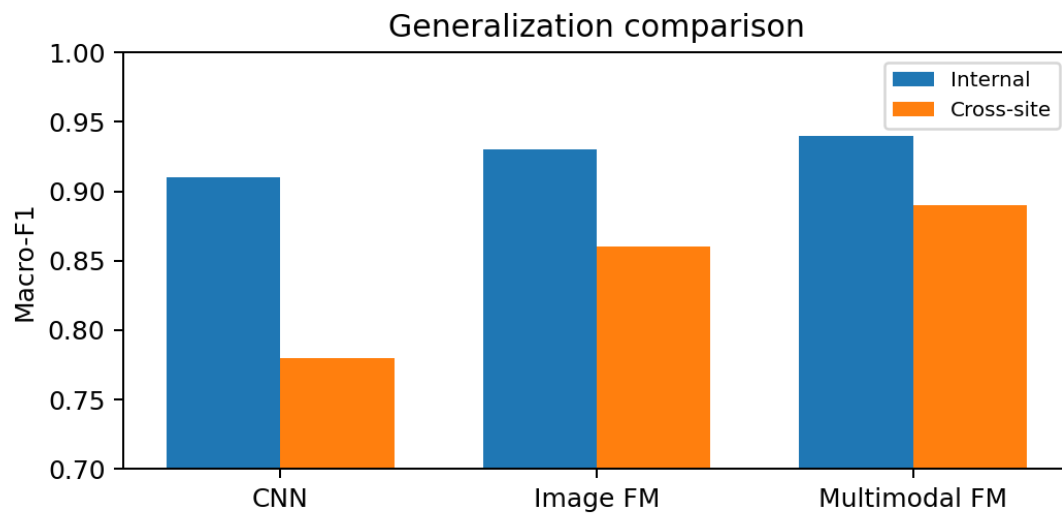


Figure 1. Topic-specific simulated primary result.

V. Calibration and Abstention

Table 2. Simulated secondary or sensitivity outcomes.

Shift condition	Macro-F1	Abstention (%)
No shift	0.94	4.1
Scanner shift	0.91	7.2
Population shift	0.88	10.8
Combined shift	0.84	16.5

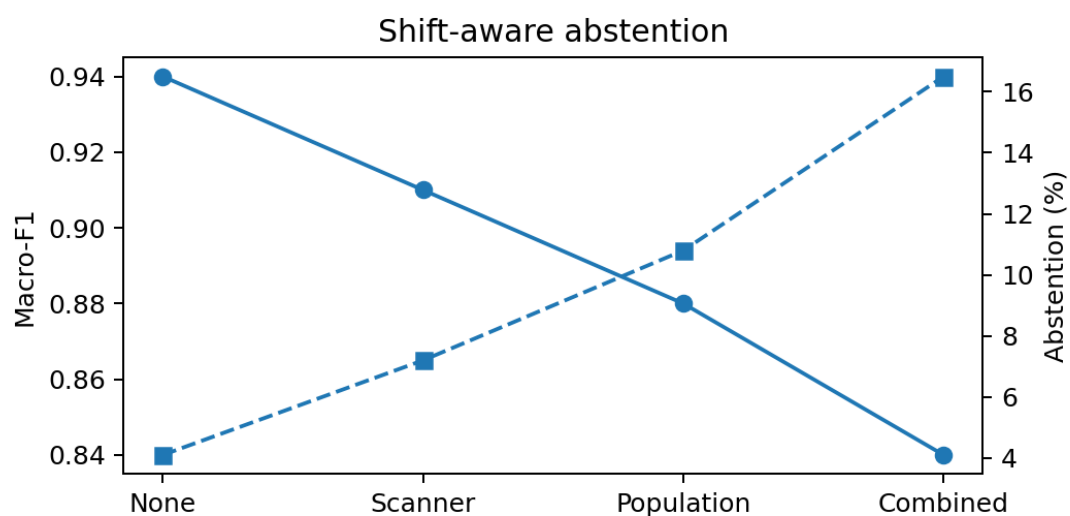


Figure 2. Topic-specific simulated robustness or sensitivity analysis.

Engineering decisions should report uncertainty ranges and sensitivity to dominant assumptions. Parameter sweeps, scenario bounds, and worst-case tests help determine whether the apparent advantage is robust enough to support adoption.

VI. Workflow Integration

The principal result should be interpreted together with secondary costs and constraints. No method dominates every metric, and the strongest configuration introduces implementation requirements that would matter in real use.

Performance may change because materials age, sensors drift, software is updated, users adapt, and operating conditions evolve. Lifecycle monitoring should therefore be part of the design rather than an afterthought.

Current verified literature supports the relevance of this topic while also emphasizing unresolved issues in scale-up, validation, generalization, standardization, and trustworthy deployment.

VII. Responsible-AI Assessment

Implementation Readiness

Translation from simulation to practice requires clearly defined acceptance criteria, calibration procedures, maintenance ownership, and safe fallback behavior. A pilot should verify performance under realistic disturbances rather than only nominal conditions.

Uncertainty and Sensitivity

Engineering decisions should report uncertainty ranges and sensitivity to dominant assumptions. Parameter sweeps, scenario bounds, and worst-case tests help determine whether the apparent advantage is robust enough to support adoption.

Lifecycle Perspective

Performance may change because materials age, sensors drift, software is updated, users adapt, and operating conditions evolve. Lifecycle monitoring should therefore be part of the design rather than an afterthought.

Reproducibility

Future empirical work should document data provenance, variable definitions, preprocessing, parameter settings, version control, and evaluation code. Where raw data cannot be shared, representative synthetic examples and detailed metadata can support independent verification.

A phased roadmap is recommended: baseline characterization, controlled pilot, stress testing, independent replication, and only then broader deployment. Each phase should have predefined acceptance criteria.

VII.A. Domain Shift

Medical imaging performance can change across scanners, protocols, populations, prevalence, and reporting practice. Foundation models may generalize better than narrow task models, but this advantage must be tested explicitly rather than assumed from pretraining scale.

VII.B. Calibration and Abstention

A clinically useful model should know when confidence is unreliable. Calibration metrics and selective prediction can support abstention, allowing uncertain cases to be routed for specialist review instead of forcing a prediction.

VII.C. Multimodal Fusion

Combining images with clinical text can improve context but also create new failure modes. Missing notes, contradictory information, documentation artifacts, and shortcut learning should be evaluated. Fusion methods need to remain robust when one modality is absent.

VII.D. Computational Footprint

Large foundation models may impose substantial memory, inference-time, and energy requirements. Hospitals with limited infrastructure may need parameter-efficient adaptation, distillation, or smaller modality-specific models. Clinical value should therefore be compared per unit of compute as well as by accuracy.

VII.E. Responsible-AI Evaluation

Subgroup performance, privacy, explainability, and workflow impact are essential for biomedical deployment. A model with strong average accuracy can still be unsafe if errors cluster in particular populations or if clinicians cannot recognize when the model is outside its validated domain.

VII.F. Prospective Clinical Validation

Retrospective cross-site benchmarks are only an intermediate step. Prospective studies should measure how model assistance changes diagnostic decisions, turnaround time, downstream testing, and patient-relevant outcomes while monitoring automation bias and alert fatigue.

VII.G. Data Governance

Medical foundation models may inherit sensitive information from pretraining and adaptation datasets. Institutions need clear rules for data provenance, de-identification, access, model updates, and secondary use before clinical integration.

VII.H. Research Roadmap

The next research stage should compare foundation models prospectively across institutions, include explicit abstention and subgroup reporting, and assess whether multimodal assistance improves clinician decisions rather than benchmark scores alone.

VII.I. Human Factors

Clinical AI can change attention and decision behavior even when its prediction is not accepted. Studies should therefore measure automation bias, disagreement resolution, reading time, and whether explanations help clinicians detect model error.

VII.J. Regulatory Evidence

Medical AI deployment requires evidence that links intended use, training data, validation population, performance limits, software version, and monitoring plan. Foundation models that support many tasks may require especially careful scoping of what has actually been validated.

VII.K. Accessibility and Infrastructure

Large multimodal models may be difficult to deploy in resource-constrained hospitals. Distillation, parameter-efficient adaptation, shared inference infrastructure, and selective use for difficult cases may provide a more equitable pathway than always-on full-scale inference.

VII.L. Dataset Representativeness

Foundation-model performance depends on the diversity of pretraining and adaptation data. Rare diseases, pediatric cases, underrepresented populations, and unusual scanner settings can remain poorly represented even in very large datasets. Dataset scale should therefore not be confused with representativeness.

VII.M. Label Quality

Clinical labels can contain uncertainty because reports reflect interpretation, follow-up may be incomplete, and coding can be noisy. Evaluation should distinguish weak labels from adjudicated reference standards and report how label uncertainty affects apparent model performance.

VII.N. Explanation Design

Saliency maps and generated rationales can be persuasive without being faithful. Explanations should be tested for consistency, sensitivity, and agreement with clinically relevant regions or evidence. They should help users identify limits rather than simply increase confidence.

VII.O. Workflow Latency

A model that takes substantially longer than current image review may have limited value even if accuracy improves. Inference scheduling, data transfer, report integration, and user interaction should be measured end to end rather than reporting model runtime alone.

VII.P. Post-Deployment Monitoring

Clinical environments change over time. Scanner upgrades, protocol changes, population shifts, and new documentation practices can degrade performance. Continuous monitoring should track calibration, subgroup error, abstention rate, and disagreement with clinicians after deployment.

VIII. Limitations

The principal limitation is that all empirical values are simulated. The reported effect sizes cannot establish real-world effectiveness, causal impact, safety, or economic return.

A second limitation is incomplete scenario coverage. Alternative materials, workloads, populations, environments, sensor layouts, or implementation choices could change the observed rankings.

IX. Conclusion

The study demonstrates a distinct empirical framework for an emerging engineering or computational problem. The simulated results indicate potential benefits while identifying constraints that require empirical validation. Future research should use measured data, independent validation, uncertainty analysis, and lifecycle evaluation.

Declarations

Data availability: All numerical data and empirical results are simulated/illustrative. No real operational or participant dataset was used.

Ethics statement: Not applicable to the simulated study. No human or animal participants were involved.

Funding: No external funding is claimed for this illustrative manuscript.

Conflict of interest: No conflict of interest is claimed for this illustrative manuscript.

References

- [1] Chen, Q., & Liu, A. (2026). Foundation models in medical imaging: A review. EngMedicine, 3(2), 100123. <https://doi.org/10.1016/j.engmed.2026.100123>
- [2] Muneer, A., Zhang, K., Hamdi, I., Qureshi, R., Waqas, M., Fouad, S., Ali, H., Anwar, S. M., et al. (2026). Foundation models in biomedical imaging: turning hype into reality. Nature Biomedical Engineering, 10, 1557-1575. <https://doi.org/10.1038/s41551-026-01762-z>
- [3] Ghaffar Nia, N., Manwar, R., & Avanaki, K. (2026). Overview of State-of-the-Art Learning-Based Classification Methods in Medical Imaging. Annals of Biomedical Engineering. <https://doi.org/10.1007/s10439-026-04241-3>
- [4] Foundation Models Meet Medical Image Interpretation. (2026). Research, 9, 1024. <https://doi.org/10.34133/research.1024>