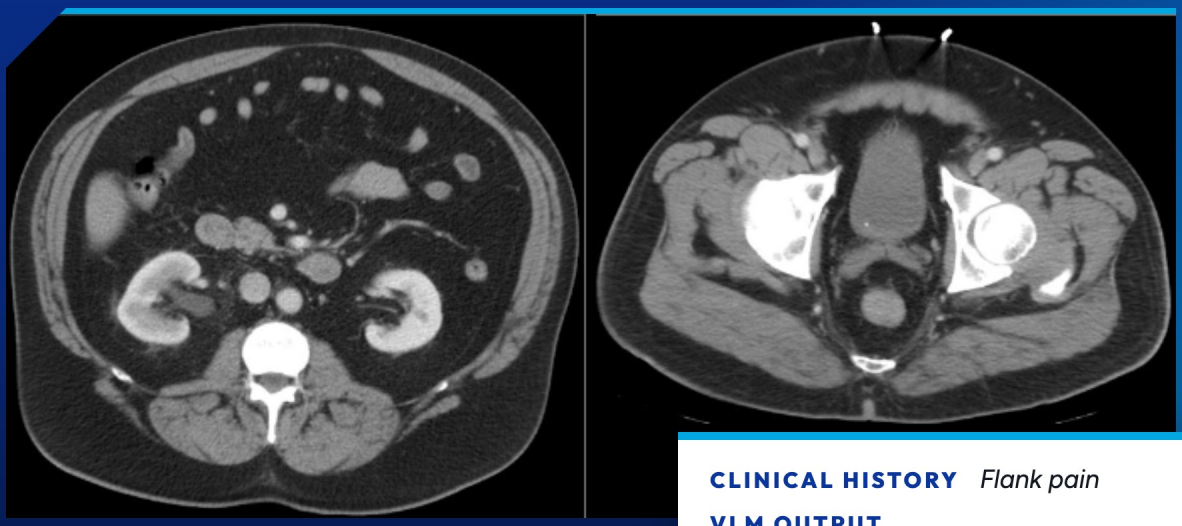
RESEARCH PAPER · CT REPORT DRAFTING

Vision-Language Models for Computed Tomography Interpretation Can Improve Report Quality and Radiologist Efficiency

Mosaic Clinical Technologies, Inc. & Cognita Imaging, Inc.

VLM OUTPUT EXAMPLE · ABDOMEN AND PELVIS CT



CLINICAL HISTORY *Flank pain*

VLM OUTPUT

- **Obstructing 2.0 mm right ureterovesical junction stone with delayed right nephrogram, mild right hydroureteronephrosis, and right perinephric and periureteral fat stranding.**

109.6%
highest rEDR: diffuse edema/ischemia, Head CT

71,707
CT exams in performance testing

27%+
CT drafting efficiency gain

RESEARCH PAPER SERIES

CT X-RAY

Vision-Language Models for Computed Tomography Interpretation Can Improve Report Quality and Radiologist Efficiency

Mosaic Clinical Technologies, Inc. & Cognita Imaging, Inc.

DRAFTED EXAMS (ALL TIME)
1.7M
CT exams drafted through Oct. 22, 2026

DRAFTED REPORTS PER DAY
12,000+
CT exams drafted per day, 30-day average through Oct. 22, 2026

EXAMS COVERED
Head CT · Cervical Spine CT · Chest CT · Abdomen and Pelvis CT

KEY FINDINGS

Highest relative enhanced detection rate (rEDR) by CT model: additional expert-confirmed findings identified by the vision-language model (VLM) compared with original radiologist reports (retrospective analysis).

109.6%
Head CT
Diffuse Edema/Ischemia

106.4%
Cervical Spine CT
Severe Stenosis

44.8%
Chest CT
Thoracic Aorta Pathology

68.1%
Abdomen and Pelvis CT
Fracture

71,707
CT exams in performance testing; 3,366 manually reviewed

71.9–87.3%
of VLM drafts rated "good" (no/minor edits), vs. 51.9–83.8% of radiologist reports

27%+
CT drafting efficiency gain (weighted, radiologists with >50 drafted exams)

01 PROBLEM-ORIENTED SUMMARY

INTRODUCTION

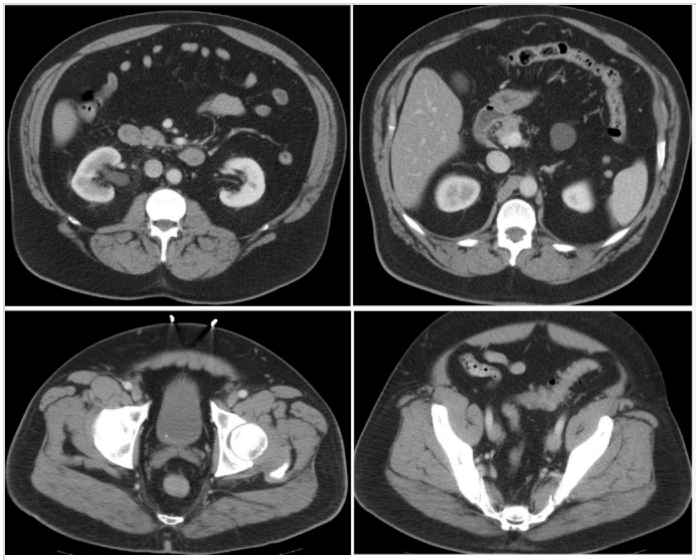
Radiology faces a growing mismatch between imaging demand and the available physician workforce, creating an urgent need for technologies that improve efficiency without compromising diagnostic quality.[1] AI-powered report drafting offers a promising solution by reducing the time radiologists spend generating reports while potentially enhancing the accuracy and completeness of the report.[2,3]

CAUTION: Investigational device. Not for sale in the United States.

Historically, radiology AI has focused on narrow detection tasks, identifying individual abnormalities such as pulmonary embolism (PE), intracranial hemorrhage, or fractures. Vision-language models (VLMs) offer the potential to move beyond this narrow paradigm.[4] Rather than generating a single binary prediction, VLMs can analyze medical images and produce multiple natural-language observations describing all relevant imaging findings.[5]

Cognita Imaging, Inc.™ (Cognita), the advanced AI lab of Mosaic Clinical Technologies, Inc.™ (Mosaic), has developed novel VLMs to assist radiologists with interpretations (Figures 1 and 2). Capabilities currently include drafting for radiographs (detailed in the accompanying radiograph research paper) and computed tomography (CT) examinations which include Head CT without contrast, Cervical Spine CT, Chest CT (with or without contrast, including CT angiography), and Abdomen and Pelvis CT (with or without contrast). The following report describes results from our research program, in which the VLMs are used under a research protocol, including performance data, user acceptance testing, and efficiency gains for CT drafting. A comprehensive lifecycle clinical AI governance process including continuous post-deployment monitoring has been established for CT drafting to ensure diligent oversight, mitigate risks and better understand the real-world effects of this new radiologist-AI collaborative technology.

Figure 1. Abdomen and Pelvis CT VLM Output Example



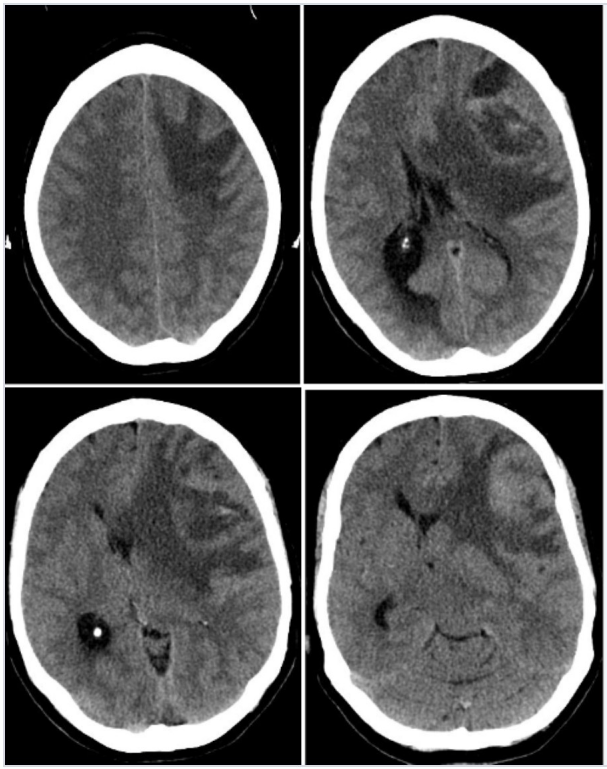
CLINICAL HISTORY

Flank pain

VLM OUTPUT

- **Obstructing 2.0 mm right ureterovesical junction stone with delayed right nephrogram, mild right hydroureteronephrosis, and right perinephric and periureteral fat stranding.**
- **2.6 x 2.8 x 2.6 cm simple fluid density lesion in the left upper quadrant, possibly duplication cyst or lymphangioma**
- **Diffuse hepatic steatosis**
- **Mild sigmoid diverticulosis**
- **Mild thoracolumbar spondylosis**

Figure 2. **Head CT VLM Output Example**



CLINICAL HISTORY

expressive aphasia

VLM OUTPUT

- **Large left frontal lobe mass measuring 4.2 x 3.4 cm.**
- **Extensive surrounding vasogenic edema.**
- **Mass effect in the left frontal region with left-to-right midline shift measuring 1.2 cm.**
- **Curvilinear hyperdensity within the mass, possibly representing hemorrhage.**
- **Effacement of left lateral ventricle.**
- **Dilatation of right lateral ventricle.**

02 CLINICAL AI GOVERNANCE

A comprehensive lifecycle clinical AI governance process is in place for CT drafting (Figure 3), which includes implementation of continuous monitoring of radiologist agreement rates with VLM output and report signing times to capture data on radiologist-AI interaction. This allows us to evaluate the real-world distribution of radiologist-AI behavior to assess model failure modes, automation bias, and outlier behavior.

Our AI continuous monitoring dashboard (Figure 4) tracks two AI-radiologist interaction metrics for each tracked high-priority pathology on drafted CT studies. The Positive Concordance rate represents the percentage of AI positive drafted cases for a specific finding that are also confirmed to be positive in the final signed radiologist reports and is a proxy for AI precision. The Negative Discordance rate represents the percentage of AI negative drafted cases for a specific finding that are found to be positive in the final signed radiologist reports and is a proxy for the AI miss rate.

Figure 3. Clinical AI Governance Process

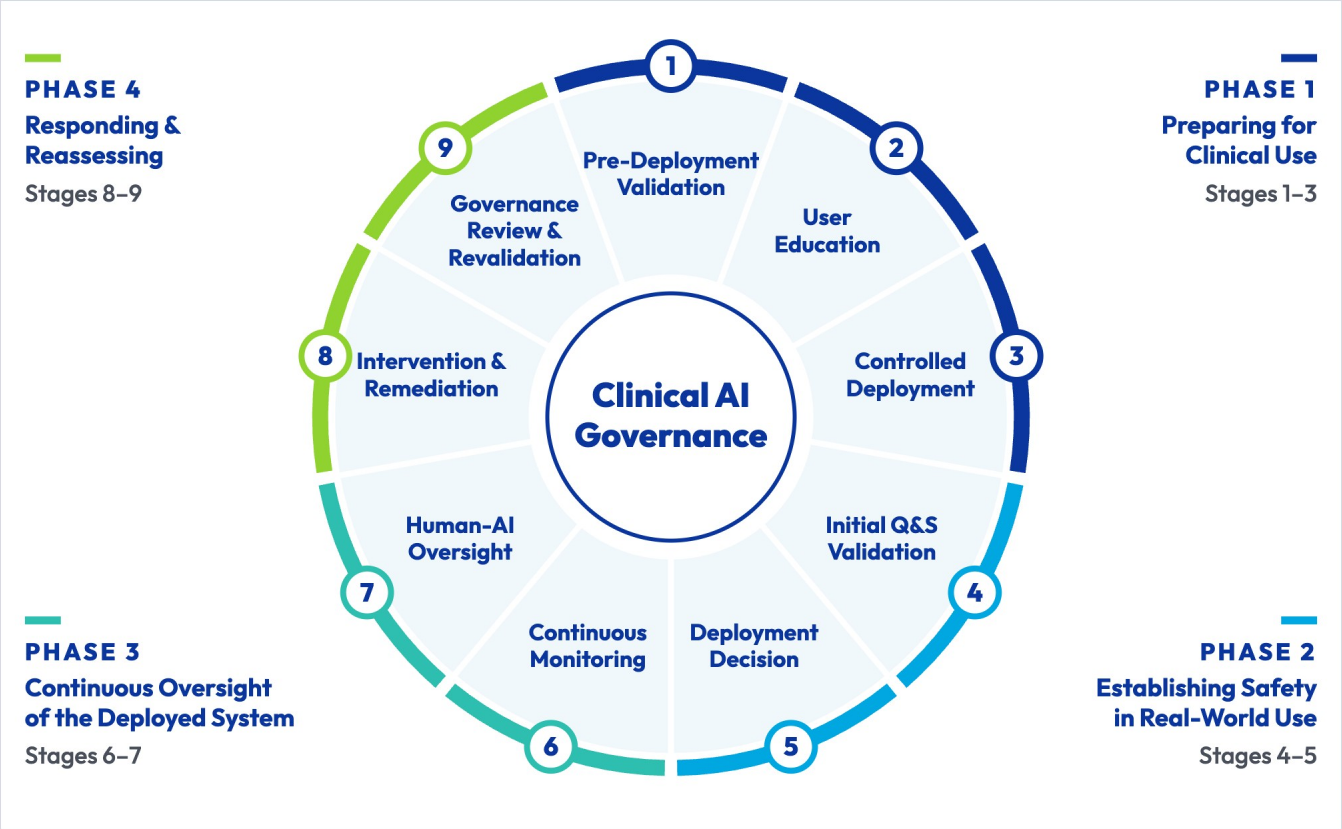
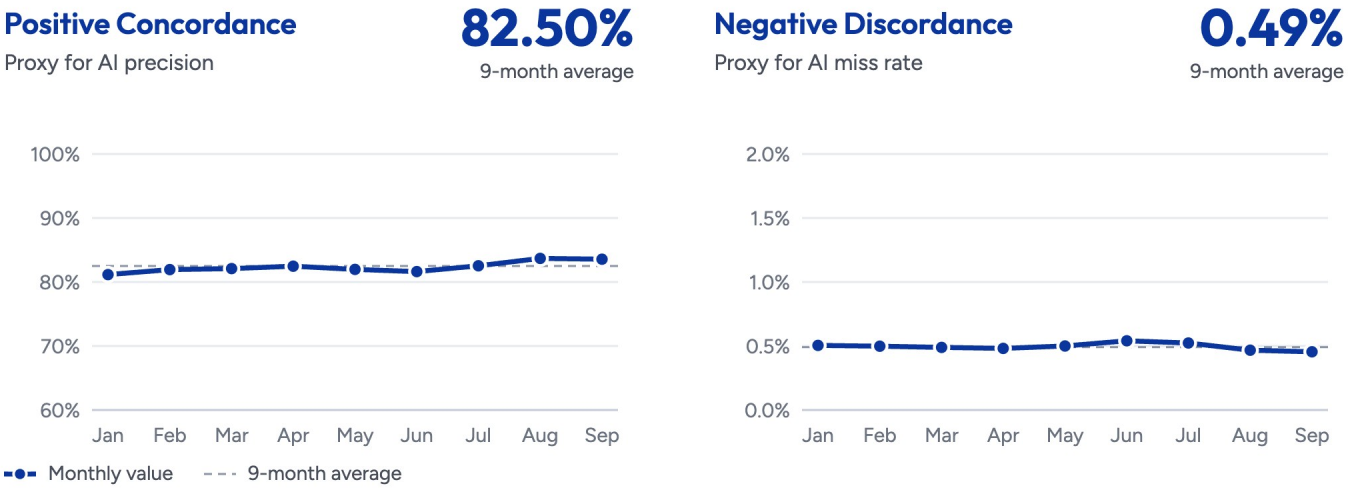


Figure 4. Head CT Continuous Monitoring Dashboard, January 1 – September 30, 2026.



Aggregated post-deployment monitoring data from 6 tracked high-priority pathologies on noncontrast Head CT studies (see Table 1 for the pathology list, excluding diffuse edema/ischemia and dural venous sinus thrombosis). Average Positive Concordance rate was 82.50%. Average Negative Discordance rate was 0.49%.

03 MODEL VALIDATION METHODS

Methods and Clinical Validation Framework for Responsible AI Implementation

To evaluate the potential clinical impact of report drafting, Mosaic developed a two-part pre-deployment clinical validation framework combining Performance Testing to measure diagnostic accuracy and opportunities for enhanced radiologist detection, and User Acceptance Testing (UAT) to assess clinical usability and editing burden.

Performance Testing: Does the VLM Detect the Full Spectrum of Disease with High Accuracy and Augment the Radiologist?

Performance testing used large multicenter, geographically diverse retrospective datasets of CT examinations processed through the corresponding VLM on patients aged 1 year and older. AI-generated outputs and outputs from the findings sections of final radiologist reports from the retrospective dataset were classified for the presence or absence of pathology using internally validated large language model (LLM) classifiers.

Multiple experienced U.S. board-certified radiologists who included a mix of subspecialists and generalists independently and blindly reviewed a natural prevalence sample of CT studies from a large retrospective dataset of CT examinations interpreted by radiologists in the Radiology Partners Clinical Services practice before AI drafting was in use. VLMs for Head CT, Cervical Spine CT, Chest CT, and Abdomen and Pelvis CT were evaluated. Specific high-priority pathologies were chosen by Mosaic for each CT exam type, as described below. We identified instances of both discordant cases, where the VLM and the original report had opposing outputs for a specific pathology, and positive concordant cases, where the VLM and the original report both registered the presence of the same specific pathology. We conducted a human-in-the-loop manual review of a randomly selected subset of both discordant cases and positive concordant cases from the full dataset to establish an expert reference standard, from which expert-adjudicated traditional model performance metrics, including prevalence, sensitivity and positive predictive value (PPV), were computed. For each examination review, the output generated by the pertinent VLM was presented alongside the output derived by an LLM from the findings section of the final signed radiologist report, with the reviewer blinded to the identity of the two positive statement lists. We further computed the augmentation metric relative enhanced detection rate (rEDR)[6], which represents the number of enhanced detection cases due to the VLM divided by the number of true positive diagnoses made by the radiologist without access to AI.

A total of 71,707 CT examinations were included for performance analysis across Head CT, Cervical Spine CT, Chest CT, and Abdomen and Pelvis CT examinations. Among these CT examinations, 3,366 exams (a mix of discordant and positive concordant cases) were manually overread by the reviewing radiologists.

Performance Testing Cohort. Retrospective CT Dataset, Patient Ages and Reviewer Panels

Exam Type	Cases Processed by VLM	Adult (Ages 18+)	Pediatric (Ages 1–17)	Cases Manually Reviewed	Reviewing Radiologists
Head CT	19,092	95.8%	4.2%	499	5
Cervical Spine CT	12,538	97.7%	2.3%	368	9
Chest CT	22,719	99.4%	0.6%	1,325	15
Abdomen and Pelvis CT	17,358	97.4%	2.6%	1,174	11
Total	**71,707**	—	—	**3,366**	—

Manually reviewed cases are a mix of discordant and positive concordant cases.

Reviewers are U.S. board-certified radiologists.

User Acceptance Testing: Is the Output Clinically Usable?

The same group of radiologist reviewers also assessed outputs along two clinically relevant dimensions at the case level. UAT was performed on a natural prevalence sample of 696 CT examinations (142 head, 158 cervical spine, 196 chest and 200 abdomen and pelvis), with reviewers blinded to whether each report was generated by the VLM or the radiologist. While performance testing evaluated individual findings, user acceptance testing evaluated each draft and radiologist report as a whole. First, reviewers categorized clinical errors as absent; present without significant downstream clinical impact; or present with potentially significant downstream impact. Second, they assessed anticipated editing burden, categorizing the statements as “good” if they could be signed with no or minimal changes, “decent” if moderate editing was required, or “problematic” if starting from scratch would be preferable.

04 SUPPORTING CLINICAL EVIDENCE

HEAD CT PERFORMANCE

A total of 19,092 cases were processed by the Head CT VLM (95.8% adult ages 18+; 4.2% pediatric ages 1-17) and there was a manual review of 499 cases (a mix of discordant and positive concordant cases) by a team of 5 U.S. board-certified radiologists. The Head CT VLM demonstrated the potential to enhance radiologist detection across all pathologies with rEDRs ranging from 5.2% for intracranial hemorrhage to 109.6% for diffuse edema/ischemia (Table 1). These potential improvements to disease detection are achieved while maintaining high positive predictive values (above 80% for 7 of the 8 conditions).

Table 1. Head CT Model Performance

Head CT Conditions	Prevalence	Sensitivity	PPV	rEDR
Intracranial Hemorrhage	8.5%	91.9%	93.3%	5.2%
Acute/Subacute Infarct	3.9%	96.7%	96.9%	14.0%
Intra-Axial Mass	1.7%	92.4%	87.4%	52.4%
Mass Effect/Herniation	5.7%	93.8%	88.3%	18.2%
Diffuse Edema/Ischemia	0.7%	100%	80.8%	109.6%
Hydrocephalus	0.4%	79.5%	89.3%	7.7%
Acute Fracture	2.7%	91.8%	89.3%	73.4%
Dural Venous Sinus Thrombosis	0.04%	71.4%	62.5%	75.0%

CERVICAL SPINE CT PERFORMANCE

A total of 12,538 cases were processed by the Cervical Spine CT VLM (97.7% adult ages 18+; 2.3% pediatric ages 1-17) and there was a manual review of 368 cases (a mix of discordant and positive concordant cases) by a team of 9 U.S. board-certified radiologists. In Table 2, the Cervical Spine CT VLM demonstrated significant enhanced detection across several pathologies (rEDRs all ≥58.6%), including low-prevalence conditions, although the false positive rate was generally higher compared to the Head CT model.

Table 2. Cervical Spine CT Model Performance

Cervical Spine CT Conditions	Prevalence	Sensitivity	PPV	rEDR
Trauma	2.4%	78.1%	61.6%	58.6%
Infection/Neoplasm	1.1%	80.5%	80.1%	70.3%
Severe Stenosis	21.2%	95.5%	100%	106.4%
Hardware Complication	0.6%	79.9%	38.9%	71.9%

CHEST CT PERFORMANCE

A total of 22,719 cases were processed by the Chest CT VLM (99.4% adult ages 18+; 0.6% pediatric ages 1-17) and there was a manual review of 1,325 cases (a mix of discordant and positive concordant cases) by a team of 15 U.S. board-certified radiologists. In Table 3, the Chest CT VLM demonstrated meaningful enhanced detection across all tested pathologies (rEDRs all ≥8.9%) with generally high PPV. The VLM provided meaningful enhanced detection for PE (19.5% rEDR), acknowledging an outlier higher false positive rate for PE for this version of the model, noted clinically to correspond to peripheral PE positive statements by the model.

Table 3. Chest CT Model Performance

Chest CT Conditions	Prevalence	Sensitivity	PPV	rEDR
Thoracic Aorta Pathology	5.0%	75.8%	83.1%	44.8%
PE	2.2%	80.5%	51.8%	19.5%
Thoracic Adenopathy	11.2%	91.3%	89.7%	43.8%
Mediastinal Pathology	45.4%	77.7%	94.3%	22.9%
Pleural Space	20.1%	91.3%	98.4%	8.9%
Lung Nodules ≥ 5 mm / Masses	22.1%	94.4%	90.3%	22.5%
Lung Findings	52.1%	93.9%	96.1%	18.6%
Thoracic Bone Pathology	12.0%	88.7%	90.1%	37.7%
Medical Devices	8.3%	98.9%	96.3%	35.9%

ABDOMEN AND PELVIS CT PERFORMANCE

A total of 17,358 cases were processed by the Abdomen and Pelvis CT VLM (97.4% adult ages 18+; 2.6% pediatric ages 1-17) and there was a manual review of 1,174 cases (a mix of discordant and positive concordant cases) by a team of 11 U.S. board-certified radiologists. The VLM demonstrated potential improvement in radiologist detection across all pathologies, most notably in vascular pathology (59.9%), lymphadenopathy and splenomegaly (53.3%), and musculoskeletal pathology (67.0%) (Table 4a). Focused evaluation of high-priority abdomen and pelvis findings including acute appendicitis, pneumoperitoneum, acute diverticulitis, pancreatic masses, and fracture further demonstrated high sensitivities of ≥82.2% (Table 4b).

Table 4a. Abdomen and Pelvis CT Model Performance

Abdomen and Pelvis CT Conditions	Prevalence	Sensitivity	PPV	rEDR
Acute Infectious, Inflammatory	14.6%	88.5%	82.2%	24.5%
Trauma, Hemorrhage	1.1%	92.8%	77.3%	20.0%
Vascular	3.8%	74.6%	84.1%	59.9%
Liver, Biliary, Pancreas	19.9%	91.4%	88.5%	26.0%
Bowel, Peritoneal	39.2%	74.6%	93.1%	11.4%
Genitourinary	26.8%	90.4%	88.0%	17.2%
Lymphadenopathy, Splenomegaly	6.9%	83.9%	97.3%	53.3%
Musculoskeletal	3.9%	95.7%	89.0%	67.0%
Lung Base	14.9%	88.1%	92.0%	24.1%

Table 4b. High-Priority Abdomen and Pelvis CT Findings

Abdomen and Pelvis CT Conditions	Prevalence	Sensitivity	PPV	rEDR
Acute Appendicitis	1.3%	82.2%	72.5%	16.0%
Pneumoperitoneum	1.9%	87.2%	93.2%	40.6%
Acute Diverticulitis	3.5%	96.2%	81.1%	37.2%
Pancreatic Masses	1.9%	87.9%	83.0%	50.4%
Fracture	2.0%	98.3%	83.5%	68.1%

UAT RESULTS

Table 5 demonstrates that the VLM output for all CT models has a clinically significant error rate comparable to or lower than that of the radiologist reports and a comparable or higher percentage of error-free reports. Moreover, the CT VLMs resulted in a higher percentage of reports requiring no or minimal editing and a lower percentage of problematic reports compared to the human radiologist reports.

Table 5. User Acceptance Testing Data: Report Quality and Clinically Significant Errors

Reader	"Good" Reports (No/Minor Edits)	"Decent" Reports (Moderate Edits)	"Problematic" Reports (Start from Scratch)	No Errors	Non-significant Errors	Clinically Significant Errors
HEAD CT · N=142						
VLM	**87.3%** (80.7 – 92.3%)	**11.3%** (6.6 – 17.7%)	**1.4%** (0.2 – 5.0%)	**82.4%** (75.1 – 88.3%)	**16.9%** (11.1 – 24.1%)	**0.7%** (0.0 – 3.9%)
Radiologist	83.8% (76.7 – 89.4%)	14.1% (8.8 – 20.9%)	2.1% (0.4 – 6.0%)	84.5% (77.5 – 90.0%)	13.4% (8.3 – 20.1%)	2.1% (0.4 – 6.0%)
CERVICAL SPINE CT · N=158						
VLM	**77.2%** (69.9 – 83.5%)	**19.0%** (13.2 – 26.0%)	**3.8%** (1.4 – 8.1%)	**83.5%** (76.8 – 89.0%)	**12.0%** (7.4 – 18.1%)	**4.4%** (1.8 – 8.9%)
Radiologist	51.9% (43.8 – 59.9%)	39.2% (31.6 – 47.3%)	8.9% (4.9 – 14.4%)	72.2% (64.5 – 79.0%)	21.5% (15.4 – 28.8%)	6.3% (3.1 – 11.3%)
CHEST CT · N=196						
VLM	**71.9%** (65.1 – 78.1%)	**25.5%** (19.6 – 32.2%)	**2.6%** (0.8 – 5.9%)	**63.8%** (56.6 – 70.5%)	**28.1%** (21.9 – 34.9%)	**8.2%** (4.7 – 12.9%)
Radiologist	54.1% (46.8 – 61.2%)	37.8% (30.9 – 44.9%)	8.2% (4.7 – 12.9%)	52.6% (45.3 – 59.7%)	39.8% (32.9 – 47.0%)	7.7% (4.3 – 12.3%)
ABDOMEN AND PELVIS CT · N=200						
VLM	**74.0%** (67.3 – 79.9%)	**23.5%** (17.8 – 30.0%)	**2.5%** (0.8 – 5.7%)	**63.0%** (55.9 – 69.7%)	**28.0%** (21.9 – 34.8%)	**9.0%** (5.4 – 13.9%)
Radiologist	62.5% (55.4 – 69.2%)	31.5% (25.1 – 38.4%)	6.0% (3.1 – 10.2%)	53.0% (45.8 – 60.1%)	38.0% (31.2 – 45.1%)	9.0% (5.4 – 13.9%)

95% Clopper-Pearson confidence intervals.

05 OPERATIONAL EFFICIENCY & WORKFLOW VALUE

PROSPECTIVE EFFICIENCY GAINS

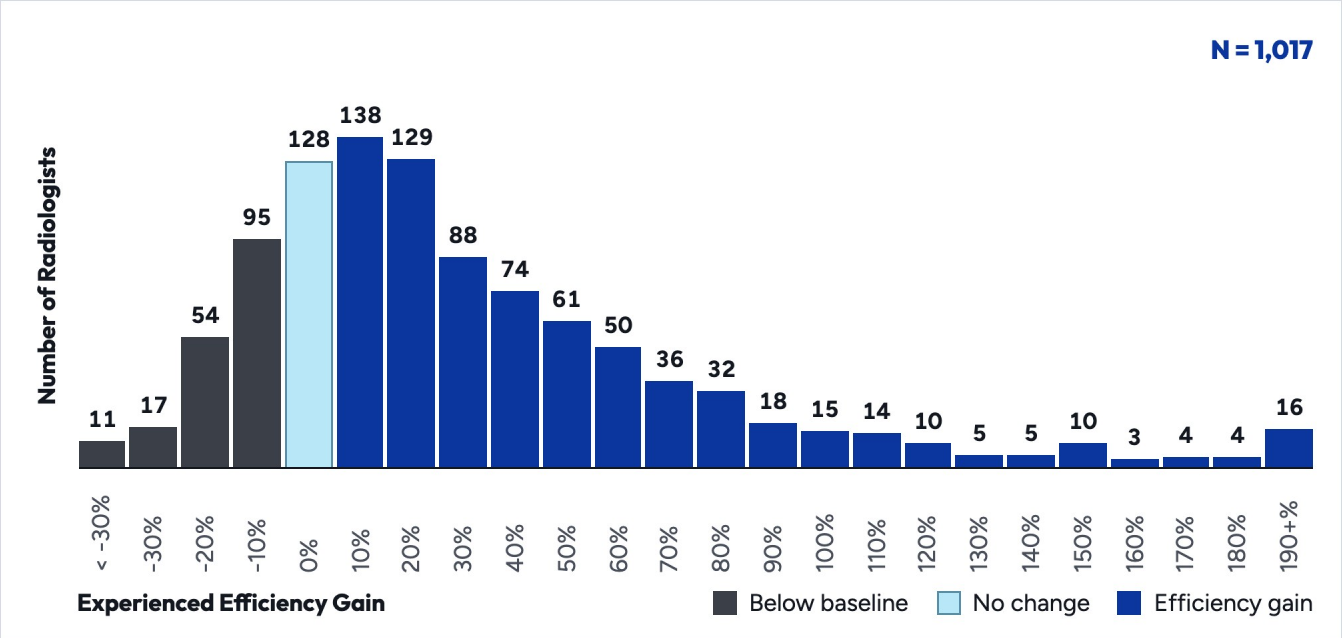
While the results above focus on retrospective evaluation, we also wanted to understand the real-world impact of radiologists working collaboratively with the VLMs. Table 6 shows the total number of examinations drafted by the models for Head CT, Chest CT, and Abdomen and Pelvis CT through Oct. 22, 2026, with overall volume-weighted CT drafting efficiency gains[†] of over 27% among participating radiologists in the Radiology Partners Clinical Services practice. Figure 5 demonstrates significant efficiency gains for approximately 77% of radiologists using drafting capabilities for both radiographic and CT examinations.

Table 6. Efficiency Impact of Drafting by Exam Type

Exam Type	Drafted Exams (All Time)	Efficiency Gain (Radiologists with >50 Drafted Exams)
Head CT	889k	+29%
Chest CT	197k	+23%
Abdomen and Pelvis CT	521k	+27%

[†]Source: Radiology Partners Clinical Transformation data. Efficiency was measured as the average time-based work units (TBWUs) completed per minute of reading time among radiologists who had read at least 50 drafted exams, compared with each radiologist's own performance before drafting was available.

Figure 5. Distribution of Mosaic Efficiency Impact, including drafted radiographs and CTs.



N = 1,017 radiologists. Each bar shows a +/-5% range around its label.

Baseline of the same radiologists reading before the implementation of the drafting solution.

06 CONCLUSIONS

Comprehensive VLMs represent a significant advancement in radiology AI, moving beyond isolated detection tasks toward clinically useful, examination-level assistance. Mosaic's ongoing research validation results demonstrate strong real-world performance across Head CT, Cervical Spine CT, Chest CT, and Abdomen and Pelvis CT, supporting the potential for VLM-powered report drafting to improve radiologist diagnostic performance, efficiency, and clinical workflow.

The opportunity is not simply better AI, but better radiology through human-AI collaboration. With encouraging clinical validation results and growing pressure on the radiology workforce, the time to begin thoughtfully integrating these capabilities into clinical practice is now. Mosaic is actively pursuing FDA regulatory authorization while continuing rigorous validation, real-world performance monitoring, and model refinement to support safe, scalable deployment.

07 REFERENCES

[1] Jing AB, Garg N, Zhang J, Brown JJ. AI solutions to the radiology workforce shortage. *npj Health Syst*. 2025;2(1):20. doi:10.1038/s44401-025-00023-6

[2] Mamdouh D, Attia M, Osama M, et al. Advancements in radiology report generation: a comprehensive analysis. *Bioengineering (Basel)*. 2025;12(7):693. doi:10.3390/bioengineering12070693

[3] Seah JCY, Tang JSN, Tran A. Drafting the future: the dawn of AI report generation in radiology. *Radiology*. 2025;316(1):e243378. doi:10.1148/radiol.243378

[4] Blankemeier L, Kumar A, Cohen JP, et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*. 2026;652(8112):1318-1328. doi:10.1038/s41586-026-10181-8

[5] Uppuluri PC, Yang CW. The emergence of foundation models in U.S. radiology: a narrative review of clinical utility, safety, and evaluation. *Acad Radiol*. 2026;33(3):797-814. doi:10.1016/j.acra.2025.10.063

[6] Larson DB, Poff JA, Krishnan S, et al. Predicting the value of radiology artificial intelligence applications: large-scale predeployment evaluation of a portfolio of models. *AJR Am J Roentgenol*. Published online 2026. doi:10.2214/AJR.25.34340

ABOUT MOSAIC CLINICAL TECHNOLOGIES, INC.

Mosaic Clinical Technologies, Inc., a wholly owned subsidiary of Radiology Partners, Inc., develops advanced technology and proprietary vision-language AI models for medical imaging, built in-house in collaboration with radiologists in the Radiology Partners Clinical Services practice. Its MosaicOS platform brings diagnostic tools, cutting-edge AI and radiology workflows together in a cloud-native operating system designed to support radiologists and the healthcare organizations they serve. Learn more at MosaicClinical.ai and connect with us on LinkedIn and YouTube.

ABOUT COGNITA IMAGING, INC.

Cognita Imaging, Inc., Mosaic Clinical Technologies, Inc.'s advanced AI lab, develops proprietary vision-language foundation models and other generative AI capabilities for radiology. Its work is focused on helping radiologists deliver more accurate, efficient care and expanding access to high-quality healthcare. Cognita seeks to elevate the performance of all radiologists to the level of the best expert subspecialists and, longer term, to empower radiologists with digital imaging information not easily perceptible to the human eye. Learn more at cognita.ai.

RSNA 2026 EVENT HUB

Download both research papers and the results summary: mosaicclinical.ai/rsna-event-hub
Contact Mosaic: mosaicclinical.ai/contact

REGULATORY COMPLIANCE NOTE

CAUTION: Investigational device. Not for sale in the United States.