DR. SERGEY MOROZOV

# WEEKLY AI INTELLIGENCE REPORT

Radiology and Medical Imaging

**Week 32 | 2 to 8 August 2026**

Dr. Sergey Morozov | drsergeymorozov.com

*Compiled from publicly available sources (indexed peer-reviewed literature, official press releases, regulatory databases). Does not constitute medical, regulatory, or financial advice.*

| 8 | 4 | 3 | 3 |
|---|---|---|---|
| Peer-reviewed papers | Industry / FDA items | Media highlights | Key opinion leaders |

SECTION 0

## EXECUTIVE SUMMARY

*Key trends, week of 2 to 8 August 2026*

### [LLM BENCHMARK] MULTIMODAL LLMS STALL BELOW CLINICAL USEFULNESS IN RADIOLOGY

Wu Q et al. (Journal of Medical Internet Research) built RadM-Bench, 720 bilingual cases across 9 subspecialties, and scored 10 multimodal LLMs 0 to 3. All stayed below 1.5. Radiologist-selected 2D key images helped every model (+19.8% to +139.2%), but volumetric input degraded all evaluable models, and performance on curated teaching cases did not transfer to routine Chinese clinical cases (O3 with images fell from 1.14 to 0.79). Translating clinical histories into English made all 10 models worse.

### [CLINICAL PREDICTION] A TRANSFUSION MODEL THAT KNOWS ITS OWN UNCERTAINTY

Li X et al. (JMIR Medical Informatics) developed MCGB, a clinically constrained gradient-boosting framework predicting transfusion need and dose in 849 patients with upper GI bleeding across 3 hospitals. AUROC 0.97, sensitivity 0.99, specificity 0.87 on a fully held-out hospital, with dose estimates carrying conformal 95% prediction intervals (coverage 0.94). Monotonic clinical constraints and quantified uncertainty answer the two standard objections to bedside ML.

### [BIAS] DATASET DIVERSITY BEATS DATASET SIZE IN GLIOBLASTOMA SEGMENTATION

Lee RS et al. (AJNR) audited four nnUNet glioblastoma segmentation models on 480 independent patients. The model trained only on white, non-Hispanic men scored lowest (Dice 0.943 FLAIR, 0.909 enhancing) and showed age and smoking-status bias; the moderate-sized but diverse BraTS 2024 model scored highest (0.996, 0.999) and beat the far larger FeTS model. Heterogeneity reduced bias even at identical training-set size.

### [OPERATIONS] LANGUAGE MODELS FIND THE MONEY LEAKING FROM ULTRASOUND NOTES

Nguyen K et al. (Journal of Medical Internet Research) used BioClinBERT (accuracy 0.97) to detect point-of-care ultrasound procedures in 559,029 OBGYN encounters, then showed a standardized documentation template cut missed-charge recapture from 10.0% to 2.4% (OR 0.22, P<.001), with 75.4% of billing codes flowing automatically from templates. AI as revenue-cycle auditor, verifying that the gain came from workflow change, not more procedures.

**[WATCHPOINT] THE WORKSTATION, NOT THE MODEL, IS WHERE THIS WEEK HAPPENS**

Read together, this week's papers relocate the frontier from accuracy to implementation. Wu Q et al. show multimodal LLMs still below clinically actionable levels once cases stop being curated. Zhang R et al. find that prompt strategy must be matched to model architecture, and that reasoning modes and parameter counts can subtract value in report classification. Moturu A et al. watch two radiologists use a pediatric whole-body MRI overlay and find it slows both down while changing their marking behavior in opposite directions, concluding that AI tools may need per-reader calibration. Even the strongest clinical result, the MCGB transfusion model of Li X et al., earns its credibility from constraints, calibration and cross-site testing rather than raw discrimination. The procurement question this week is not what a model scores, but what happens to the humans, the workflow and the billing file when it arrives.

SECTION 1

# PEER-REVIEWED PAPERS

*Source: PubMed | Verified indexing | No preprints | 2 to 8 August 2026*

## Prehospital Injury Severity Estimate (PHISE) matches in-hospital trauma scores when embedded in AI models.

*Sigle M, et al.* | ***NPJ digital medicine*** | *2026-08-07*

PHISE, a scene-based prehospital injury severity score built by mapping AIS-coded injury descriptions from the US National Trauma Data Bank to eight body regions and four severity levels, using LLM-assisted inference to flag which injuries need imaging and which can be assessed clinically. The goal is an anatomical severity component that works before any scanner is reached, and that can be embedded into machine learning models with other prehospitally available variables.

**Key metrics:** The abstract reports the comparisons qualitatively, without numeric values: PHISE showed strong correlation and calibration with ISS and NISS in the NTDB cohort but lower predictive performance for in-hospital mortality and transfusion need. In external validation on the multinational TraumaRegister DGU it underperformed ISS/NISS for mortality, matched them for transfusion prediction, and the gap narrowed further when PHISE was embedded in ML models with prehospital variables.

**Clinical relevance:** Trauma triage decisions are made before imaging exists, yet the reference severity scores all depend on it. An imaging-independent anatomical score that is deliberately designed as an AI model input, and externally validated on a registry from a different trauma system, is a credible route to earlier, data-driven trauma stratification and smarter activation of imaging and transfusion resources.

**PMID** 42562846 **DOI** 10.1038/s41746-026-03074-7

## Prediction of Blood Transfusion Need and Dose in Patients With Upper Gastrointestinal Bleeding: Retrospective Multicenter Prediction Model Study.

*Li X, et al.* | ***JMIR medical informatics*** | *2026-08-04*

A two-stage, clinically constrained gradient boosting framework (MCGB) that first predicts whether a patient with endoscopically confirmed upper gastrointestinal bleeding will need transfusion, then estimates the dose with conformal 95% prediction intervals. Developed on a retrospective multicenter cohort of 849 adults from 3 hospitals in Chongqing (2019 to 2025), with monotonic clinical constraints, cross-site hold-out testing and a bedside GUI prototype.

**Key metrics:** AUROC 0.97 and AUPRC 0.91 for transfusion need; at the 0.50 threshold, sensitivity 0.99, specificity 0.87, F1 0.85. Dose prediction among transfused patients reached R2 0.95 with mean absolute error 0.04, and 95% prediction-interval coverage of 0.94. Evaluation used one full hospital held out as an independent test site, with hospital-wise alternating external testing as a robustness check.

**Clinical relevance:** Transfusion thresholds in UGI bleeding are contested, and fixed hemoglobin cutoffs ignore the individual patient. A calibrated model that quantifies its own uncertainty, respects clinical monotonicity, and was tested across sites addresses the two chronic complaints about ML decision support, opacity and overconfidence, and has direct implications for blood-bank planning as well as bedside decisions.

**PMID** 42550984 **DOI** 10.2196/83889

## Evaluating Sociodemographic Biases in Artificial Intelligence-Based Glioblastoma Response Assessment Algorithms.

*Lee RS, et al.* | ***AJNR. American journal of neuroradiology*** | *2026-08-03*

A bias audit of four nnUNet glioblastoma MRI segmentation models whose training sets differ in size and demographic composition: the FeTS postoperative model (over 10,000 examinations, multinational), the BraTS 2024 postoperative glioma model (over 2,000 examinations, North American), and two small single-institution models (over 200 examinations each), one demographically homogeneous, one heterogeneous. All were evaluated on an independent, manually corrected set of 480 prospectively collected patients.

**Key metrics:** The model trained exclusively on white, non-Hispanic men had the lowest Dice scores (0.943 FLAIR, 0.909 enhancing tumor) and showed biases by age and smoking status. The BraTS model had the highest Dice scores (0.996 FLAIR, 0.999 enhancing) and the least bias, outperforming the FeTS model despite FeTS having the largest training set. Sociodemographic effects were analyzed with beta regression.

**Clinical relevance:** Two findings matter for anyone assembling training data: demographic heterogeneity reduced bias even without increasing dataset size, and a moderate, diverse cohort beat a much larger one. Bias in segmentation models turns out to be low overall but concentrated where training data are homogeneous, which is an argument for auditing composition, not just volume, in procurement and validation.

**PMID** 41663205 **DOI** 10.3174/ajnr.A9217

## Machine Learning to Identify Point-of-Care Ultrasound and Evaluate Standardized Documentation: Retrospective Operational Cohort Study.

*Nguyen K, et al.* | ***Journal of medical Internet research*** | *2026-08-07*

An operational study at a large academic center that used ML (LightGBM and BioClinBERT) to detect point-of-care ultrasound procedures buried in free-text OBGYN notes, then measured what a standardized procedure documentation (ProcDoc) template did to billing capture. The analysis covered 559,029 encounters from 109,776 patients across 11 clinic sites from 2018 to 2024, with the template introduced in February 2023.

**Key metrics:** BioClinBERT reached accuracy 0.97 (F1 0.55 to 0.63) for identifying documented and missed procedures. ProcDoc adoption hit 75.1% within 12 months. Missed-charge recapture fell from 10.0% before the intervention to 2.4% after (odds ratio 0.22, 95% CI 0.17 to 0.30, P<.001), with overall POCUS billing up 0.6%. 1,812 of 2,404 postintervention CPT codes (75.4%) came from the templates.

**Clinical relevance:** This is AI as an auditing instrument rather than a diagnostic one: the models quantified revenue leakage from undocumented POCUS and then verified that the fix was workflow change, not volume growth. For imaging leaders the transferable lesson is the pairing, a language model to measure the problem in free text plus a structured documentation intervention to close it.

**PMID** 42566781 **DOI** 10.2196/84454

## A Bilingual Benchmark for Evaluating Diagnostic Performance of Multimodal Large Language Models in Radiology (RadM-Bench): Evaluation Development and Validation.

*Wu Q, et al.* | ***Journal of medical Internet research*** | *2026-08-07*

RadM-Bench, a bilingual radiology benchmark of 720 cases across 9 subspecialties: 360 English public teaching cases enriched in rare diseases and 360 Chinese routine clinical cases. Ten multimodal LLMs (4 proprietary including GPT-4o, O3 and Gemini variants; 6 open-source) were tested with history alone, radiologist-selected 2D key images, and volumetric input at 2 and 10 frames per second, scored 0 to 3 by two blinded board-certified radiologists.

**Key metrics:** Mean scores stayed below 1.5 of 3 for all 10 models on both datasets. Adding 2D key images improved every model (+19.8% to +139.2%), but volumetric input at fps=10 degraded all 8 evaluable models versus the 2D baseline (-5.3% to -31.4%). Proprietary models declined from teaching to clinical cases (O3 with images: 1.14 to 0.79) while Chinese-centric open-source models improved. Translating Chinese histories into English lowered scores in all 10 models, significantly in 9.

**Clinical relevance:** The benchmark isolates exactly the gaps that clinical deployment would hit: models cannot yet use 3D data, performance on curated teaching material overstates performance on routine cases, and the rare-disease advantage seen on teaching sets reverses in the clinic. The language finding cuts against the assumption that English input is always optimal. A useful reference before any multimodal LLM pilot in reporting or triage.

**PMID** 42566748 **DOI** 10.2196/92183

### Large Language Models for Classifying Usual Interstitial Pneumonia from Radiology Reports: Native Reasoning Versus Structured Prompting.

*Zhang R, et al.* | ***Journal of imaging informatics in medicine*** | *2026-08-05*

A systematic comparison of 10 open-source LLMs from the Llama, Qwen and Gemma families (8B to 405B parameters) for classifying usual interstitial pneumonia patterns from 270 HRCT reports against Fleischner criteria, with expert consensus of two senior thoracic radiologists as reference. Three prompting strategies were crossed with model architecture, and four models with native reasoning capability were additionally run in thinking mode.

**Key metrics:** The best configuration reached Cohen's kappa 0.70 and 82% four-class accuracy. Structured reasoning prompting improved all Llama models but degraded every model with native reasoning capability (Qwen 3.5, Gemma 4). Thinking mode hurt performance on criteria-based prompts and never produced the best configuration. Llama 3.1 405B offered no advantage over Llama 3.3 70B, and the 397B Qwen underperformed the dense 27B model.

**Clinical relevance:** For report-mining pipelines the operational findings are that prompt design must be matched to architecture, that reasoning modes can actively hurt rule-based classification, and that parameter count is a poor purchasing criterion. Groups labeling reports for model training can get state-of-the-art results from mid-size open models, provided they benchmark the prompt-architecture pairing rather than assume newer and bigger is better.

**PMID** 42557432 **DOI** 10.1007/s10278-026-02178-6

### Automatic Patient Eligibility for Photon-counting CT Using Discriminative and Generative AI Models in Neuroradiology.

*Martin-Noguerol T, et al.* | ***Academic radiology*** | *2026-08-05*

A proof-of-concept NLP router that reads Spanish-language neuroradiology CT requests and decides whether the study needs a full photon-counting CT protocol or can go to a conventional energy-integrating detector scanner. 800 requests (2012 to 2025) were labeled by two neuroradiologists (Cohen's kappa 0.74); fine-tuned discriminative transformers were compared against zero-shot generative models (ChatGPT 5.2, Gemini 3) under five-fold cross-validation.

**Key metrics:** Biomedical RoBERTa led with accuracy 0.925, precision 0.906, recall 0.967, specificity, F1 0.935 and AUC 0.920, not significantly different from Spanish BERT but significantly better than Llama-3.1-8B, Mistral-7B and both generative LLMs, which performed worst overall.

**Clinical relevance:** PCCT capacity is scarce and there are no guidelines for when it should be preferred, so an automatic eligibility router addresses a real resource-allocation gap. The result also lands a broader point for radiology IT: small fine-tuned discriminative models still beat general-purpose generative LLMs on well-defined classification, at a fraction of the cost and governance burden.

**PMID** 42557204 **DOI** 10.1016/j.acra.2026.07.026

### Human-AI Interaction With AI-Assisted Tumor Overlays in Pediatric Whole-Body Magnetic Resonance Imaging: Exploratory Reader Study.

*Moturu A, et al.* | ***JMIR human factors*** | *2026-08-07*

An exploratory human-AI interaction study of a patch-based segmentation overlay that highlights tumor-like regions on pediatric surveillance whole-body MRI in Li-Fraumeni syndrome, trained on augmented 2D slices from 675 wbMRI volumes. Two radiologists read cases with and without the overlay while evaluation time, candidate-lesion marking, follow-up recommendations and perceived workload were recorded.

**Key metrics:** The abstract reports directional results without numeric values: average evaluation time per case increased with the AI tool for both readers; one reader marked more candidate lesion locations with the tool and the other fewer; one reported lower mental demand and both reported lower stress with the tool. Interrater variability was evident throughout.

**Clinical relevance:** A candid data point on what AI assistance actually does at the workstation: it changed behavior in reader-specific directions and slowed both readers down, while reducing stress. The authors' conclusion that AI tools may need personalized calibration per radiologist challenges the one-model-fits-all deployment pattern, and the pediatric rare-disease surveillance setting is precisely where a second pair of eyes has the highest stakes.

**PMID** 42567688 **DOI** 10.2196/81066

SECTION 2

## INDUSTRY & REGULATION

*Sources: Official registers | Press releases | FDA databases*

### [MARKET] GE HealthCare launches Invenia ABUS Prime and FDA-cleared ABUS StreamVue viewer

*GE HealthCare | 3 August 2026*

GE HealthCare announced the launch of Invenia ABUS Prime, its next automated breast ultrasound system for supplemental screening in dense breasts, together with ABUS StreamVue, a zero-footprint browser-based enterprise viewer that recently received FDA 510(k) clearance and CE marking. ABUS Prime carries AI-based Scan Quality Assessment, Auto Nipple Detection and QVCAD reading support; per company materials the AI-assisted review reaches 93% sensitivity and can cut interpretation times by up to 33%, with Fast Scan reducing acquisition times by up to 40%. StreamVue, built with Candelis, targets remote reading and multi-site standardization; one of the first US Invenia ABUS Prime installations is at BeSound in Los Angeles.

Source: GE HealthCare

### [REGULATION] Cortechs.ai expands EU MDR CE marking across its quantitative imaging platform

*Cortechs.ai | 4 August 2026*

Cortechs.ai announced expanded CE marking under EU MDR for its latest products: Version 5 of NeuroQuant (including NeuroQuant MS and NeuroQuant ARIA), NeuroQuant Brain Tumor, and OnQ Prostate. The certification enables commercialization of the current generation of its automated brain volumetry, brain tumor monitoring and RSI-based prostate imaging tools across the European Economic Area. The company frames the milestone as the foundation of its European go-to-market, following more than a year of regulatory work under MDR.

Source: PR Newswire

### [PARTNERSHIP] 4DMedical invests $3.4M in RevealDx and takes global distribution of RevealAI-Lung

*4DMedical | 3 August 2026*

4DMedical announced a strategic investment of $3.4M in Seattle-based RevealDx and a distribution agreement covering the US, Europe, Australia and New Zealand for RevealAI-Lung, an AI CADx tool that characterizes pulmonary nodules with a Malignancy Similarity Index (mSI). RevealAI-Lung is FDA cleared, EU MDR certified and TGA approved, validated in more than 1,500 patients, and reimbursable in the US under CPT codes 0721T and 0722T. 4DMedical will embed it into the contextflow chest CT platform it recently acquired and pair it with its CT:VQ functional lung imaging, assembling a single cardiopulmonary ecosystem spanning detection, risk stratification and function.

Source: 4DMedical

### [MARKET] GE HealthCare introduces LOGIQ e family of AI-enabled compact ultrasound systems

*GE HealthCare | 4 August 2026*

GE HealthCare introduced the LOGIQ e Xi and LOGIQ e Si, laptop-format ultrasound systems combining console-level performance with AI features for point-of-care and multi-specialty use, including Nerveblox anatomical landmark labeling for nerve blocks, MSK Go structure identification, Automated Function Imaging for cardiac assessment, and SonoBiometry fetal measurements. Workflow tools such as Scan Assistant and Voice Control reduce manual interactions, and the Xi supports Verisound cloud fleet management and Vscan Air probe compatibility. The launch extends the AI-enabled portable imaging segment across emergency, critical care, MSK and anesthesiology settings.

Source: MPO Magazine

SECTION 3

## MEDIA HIGHLIGHTS

*AuntMinnie | Radiology Business | The Imaging Wire | Diagnostic Imaging | ITN*

### Can LLMs fill the communication gap with referring clinicians?

***AuntMinnie.com*** | *4 August 2026*

Coverage of a UCSF study published 4 August in Radiology showing that LLMs can generate more comprehensive and clinically useful imaging indications than referring clinicians provide. Drawing on 28,313 patients, 740,867 clinical notes and 77,626 examinations, the team benchmarked proprietary and open-source models; in a 20-radiologist reader study, Claude 3.5 Sonnet took top comprehensiveness ratings 37.14% of the time and top factuality 68.05%, versus 6.64% and 50% for referring clinicians, with the open-source Qwen 2.5-7B close behind. An accompanying editorial stresses transparency, monitoring and human oversight before deployment.

Link: https://www.auntminnie.com/imaging-informatics/artificial-intelligence/article/1...

### Konica Minolta expands Exa Enterprise platform with cloud teleradiology and AI reporting partners

***Radiology Business*** | *3 August 2026*

A feature on Konica Minolta Medical Imaging USA's expansion of its Exa Enterprise PACS into a cloud-based teleradiology platform assembled from partners rather than in-house builds: NewVue workflow orchestration, RADPAIR generative AI reporting, AWS Health Imaging cloud infrastructure, and Apollo VNA. The browser-based platform is priced as a single per-study fee and scales from single-physician practices to 500-bed hospital systems, illustrating the platformization trend in which PACS vendors integrate third-party AI rather than develop it.

Link: https://radiologybusiness.com/topics/health-it/enterprise-imaging/konica-minolta...

### Week in Review: report turnaround times keep climbing

***AuntMinnie.com*** | *7 August 2026*

AuntMinnie's weekly editor's note highlights Harvey L. Neiman Health Policy Institute findings that radiology report turnaround times for Medicare patients continued to rise in 2024, though more slowly than in 2023, against the backdrop of the radiologist shortage and growing volumes. It also flags a UMass Memorial Health pilot that routed urgent primary care patients directly to radiology instead of the emergency department, cutting time from arrival to imaging results by nearly 50%, a workflow result relevant to the capacity argument that drives radiology AI adoption.

Link: https://www.auntminnie.com/editors-note/article/15831989/week-in-review-imaging-...

SECTION 4

## PROFESSIONAL SOCIETIES

*SIIM | ACR | RCR | Official publications and event coverage*

### RSNA launches the 2026 Knee Abnormality Detection AI Challenge

*5 August 2026*

RSNA issued the official press release for its 2026 AI Challenge, the first to focus on musculoskeletal MRI and the first to combine images with radiology report text. The training set spans more than 5,000 knee MRI exams with reports in 12 languages from 16 sites across five continents, annotated by expert radiologists for the evaluation set. The Kaggle-hosted competition runs through 22 October, with $77,000 in prizes including a first-time award for the most efficient models; winners will be recognized at RSNA 2026 in Chicago (29 November to 3 December).

Source: https://www.rsna.org/media/press/2026/2669

SECTION 5

## KEY OPINION LEADERS

*LinkedIn | Public statements | Week of 2 to 8 August 2026*

### Curt Langlotz

*Professor of Radiology, Medicine and Biomedical Data Science, Stanford University School of Medicine*

Amplified (7 August) the publication of MedVAL in npj Digital Medicine, a Stanford paper on which he is a coauthor: Toward expert-level medical text validation with language models, published 4 August. The work addresses one of the practical bottlenecks of generative AI in radiology, verifying whether AI-generated medical text is factually consistent with its source, and proposes language models themselves as scalable validators. Langlotz also posted this week about meeting the Raidium team in Palo Alto, revisiting his 2017 line that AI will not replace radiologists, but radiologists who use AI will replace those who do not.

Source: LinkedIn

### Woojin Kim

*Chief Strategy Officer and CMIO at HOPPR, CMO of the ACR Data Science Institute, MSK radiologist*

Posted on 6 August that he will attend EuSoMII 2026 in Heraklion, Crete (9 to 10 October) representing the ACR Data Science Institute, sharing the ACR DSI campaign on how the institute is shaping the future of radiology AI. Earlier in the week (2 August) he described giving a talk on entrepreneurship and AI in radiology in Geraldine McGinty's Healthcare Innovation and Transformation course at Weill Cornell, a signal of how radiology AI leadership is entering executive healthcare education.

Source: LinkedIn

### Amine Korchi

*Radiologist and healthtech entrepreneur, Geneva; co-author of the HUG neuroradiology VLM review*

Posted on 7 August about a newly published systematic review, led with the neuroradiology team of Geneva University Hospitals (HUG), on vision-language foundation models for 3D neuroradiological interpretation. His summary of the evidence: domain-specific neuroimaging VLMs consistently outperform general-purpose multimodal models and are bringing AI-assisted report drafting closer to clinical reality, but they are not yet ready for routine adoption. Remaining gaps he lists include too many report corrections for meaningful workflow gains, heavy compute demands, the need for clinically meaningful evaluation metrics, hallucinations and missed findings, workflow integration and regulatory readiness. For AI builders he frames these limitations as the to-do list on which the next generation of radiology AI companies will be built.

Source: LinkedIn

**QUALITY ASSURANCE NOTE**

*Compiled by Dr. Sergey Morozov from publicly available sources: peer-reviewed papers indexed in PubMed (PMIDs and DOIs verified, no preprints or arXiv), official press releases, regulatory databases and specialist media. It does not constitute medical, regulatory or financial advice. All papers have publication dates verified within 2 to 8 August 2026. Paper selection is deterministic (pinned PubMed query, Q1/flagship journal allow-list, exclusion of pure reviews [meta-analyses kept], radiomics-only and interventional-radiology work, then a transparent ranking score). Industry items come from official press releases / regulatory notices; KOL items from public LinkedIn activity within the window.*