This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Auditing Surgical-AI Models Under a Fixed Annotation Budget: Unbiased Performance and Confirmed Failure Modes

Yann Sakref[1⋆], Chloe Nobuhara[1], Jeffrey K. Jopling[2], and Lisa M. Knowlton[1]

[1] Stanford University, Department of Surgery, Stanford, CA, USA
[2] Johns Hopkins University, Department of Surgery, Baltimore, MD, USA

Abstract. Performance of artificial intelligence (AI) models may degrade once deployed to a new setting with a different dataset and case mix than its training data. In surgery, where these models increasingly inform decisions with direct impact on patient outcomes, measuring that degradation and locating where a model fails is essential. We propose an approach for evaluating AI models for surgical video datasets under a fixed annotation budget, without requiring pre-existing labels, retraining of the model, or trusting its self-reported confidence. The method dedicates part of the budget to a uniform-random cohort, which gives an unbiased failure rate that stays valid under miscalibration; trains an embedding predictor on that cohort to locate candidate error regions; and spends the rest confirming each region on disjoint labels. In simulated audits across five open dataset and model pairs spanning surgical phase recognition, organ segmentation, and visual question answering, the uniform cohort holds a 3–5 percentage-point standard error with 0.96–0.99 coverage of the true error rate, while confidence-sampled estimators miss by tens of points on miscalibrated models. We also show that confident-wrong failures are 12–20% of the error mass and do not shrink as the model improves, so the model's confidence cannot be used to find them. Flagged failure regions are confirmed at false-positive rates of 0.05–0.14 against a nominal 5% target. The audit returns an unbiased performance headline with its interval, a verdict on whether the model's confidence tracks its errors, and a map of confirmed failures categorized by location and by whether they are silent or loud.

Keywords: Post-deployment auditing · Surgical AI · Label-efficient evaluation · Calibration · Failure discovery · Post-market surveillance

1 Introduction

Implementation of surgery-specific artificial intelligence (AI) models are an active area of translational research, though most reported applications at the proof-of-concept stage [19,21,33,23,26]. Surgical AI systems are increasingly built for tasks with direct impact on patient outcomes, from recognizing operative

⋆ Corresponding author: ysakref@stanford.edu

phases to delineating critical anatomy [9,20,14,18]. However, these models do not always perform with the same level of accuracy when deployed in new settings: on a different hospital's camera hardware, sites with different surgical workflows, or case mixes [34,13,15]. Regulatory approval for these AI systems does not account for the risks of site-specific failures, as it is based on available data rather than the conditions of local deployment [35,6,17]. Post-deployment model performance assessment remains the responsibility of the new site that acquired the model. This assessment, however, is constrained by costly and time-consuming expert annotation to describe frame-level anatomy, phase boundaries, or visual failure modes. Expert review of surgical video is resource-intensive, thus an audit must expend these resources judiciously.

To evaluate a surgical AI model, a site needs two things from its limited budget: a reliable performance measure (the headline), which informs whether to keep, limit, or retrain the model, and a failure map, which bounds safe clinical use and guides communication with the vendor. A single cohort struggles to deliver both: a representative sample surfaces rare failures only inefficiently, and a sample enriched with known failures biases the performance estimate. For the headline, prediction-powered inference combines a small labeled sample with the model's predictions on the unlabeled pool [3]. Those labels can be drawn in different ways, uniformly at random or, in active variants, by using the model's confidence to choose which inputs to label [36,16]; either way the method returns a headline but no failure map. For failure discovery, slice-discovery, out-of-distribution methods [8,7,30] and drift monitoring [24,13,25] point to suspect inputs but attach no validated performance figure to them, and active learning [28,27,4] spends labels to improve a model but not to audit a frozen one. A site could, in principle, run one method from each family in turn, but their costs add rather than share: labels bought for an enriched discovery cohort cannot feed an unbiased headline, and confirming a discovered region demands fresh annotation on top. An audit under a fixed budget must instead optimize the same labels to serve both objectives, while also treating the model's confidence used for discovery and active-sampling methods as something to test. To our knowledge, no existing method serves both objectives under one shared budget.

Three properties make combining both auditing aspects especially hard. First, a model's own confidence is a fragile guide to choose data to label: enrichment must guess which frames will fail, and the only guess available before spending a label is the model's confidence, which may be miscalibrated [10,16]. Second, an important fraction of errors are silent, where the model is confidently wrong. As these errors carry the same high confidence as correct predictions, they are invisible to any confidence-based safety net during deployment, hence are dangerous as most likely to be acted on. Third, discovering failure regions in a pool and then testing them on the same labels is circular, and manufactures failures that do not exist.

We address all three with an auditing protocol for surgical AI models at a single site. The headline uses a uniform-random cohort, unbiased by construction and robust to the model's miscalibration. Failure discovery uses an embedding

predictor of error that is blind to the model's confidence, the only route that reaches the silent stratum. Discovered regions are confirmed on a disjoint cohort under a multiplicity correction, which restores post-selection validity, and the output is a relocatable two-axis (location × silent/loud) failure map. Our audit builds the second objective on the first: the one uniform cohort that yields the headline also trains the failure predictor and sets the baseline against which candidate regions are confirmed. We simulate this protocol on fully labeled surgical benchmarks (Cholec80 for phase recognition [31], CholecSeg8k for segmentation [12], and PitVQA for visual question answering (VQA) [11]) and evaluate it under realistic budget constraints. These experiments are single-site simulations on public datasets: a proof of concept establishing the feasibility and statistical validity of the protocol, not a measurement of performance drift at unseen centers. This approach not only quantifies each component of the auditing process but also provides a standard framework for comparing different estimators, discovery signals, and confirmation methods within the constraints of a fixed budget.

2 Method

The inputs of our audit are a frozen model M that emits a per-input prediction and, where exposed, a confidence score; an unlabeled site pool P of X inputs; an annotation budget $Y \ll X$; and an embedding of each input, taken from M's penultimate features where available and otherwise from a generic backbone such as DINOv2 or DINOv3 [22,29]. Per-input correctness is $c_i \in \{0, 1\}$ for the binary tasks (exact match for VQA) and $c_i \in [0, 1]$ (per-frame mean intersection-over-union, mIoU) for segmentation.

Audit protocol (Fig. 1). (S1) Track A: draw a uniform-random cohort of size n_A without replacement, reveal labels, and compute the headline $\bar{e}_A$ with an interval; this cohort contains silent failures at their true rate. (S2) On Track A, verify whether confidence separates wrong from right and in which direction; this governs only whether the silent/loud axis is later reported. (S3) Embed every input, L2-normalised, and record the prediction and confidence. (S4) Train an error predictor on Track A: a plain classifier (a linear logistic probe on the embedding) that learns what failed frames look like. It never sees confidence, so it can find failures whether or not the model was sure of them. (S5) Score the whole pool with this predictor and group the high-scoring inputs into candidate regions with a count-free method (default: a Gaussian mixture selected by BIC); these are hypotheses, not yet findings. (S6) Track B: for each region draw $n_{\min}$ fresh labels inside it, disjoint from Track A. (S7) Test one-sided that each region's rate exceeds $\bar{e}_A$ (binomial for binary correctness, bootstrap for mIoU) under a multiplicity correction; keep those that pass. On each confirmed region, we use the models' confidence to tag its error fraction as silent or loud. (S8) Return the headline, the confirmed modes, the failure map as embedding coordinates plus the trained predictor, the reusable cohort $A \cup B$, and the confidence verdict.

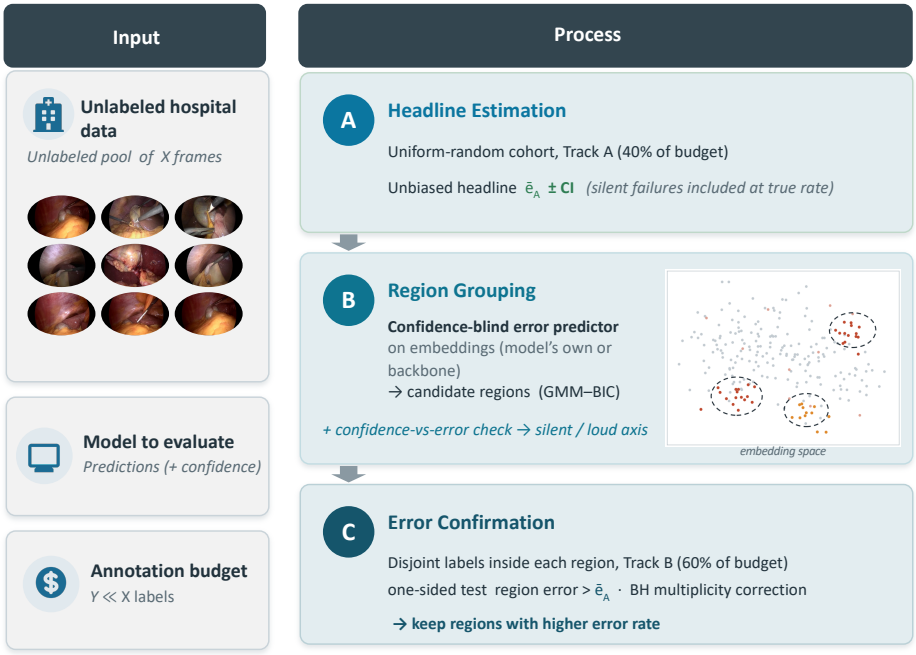


Fig. 1. Overview of the audit. Inputs (left): the unlabeled deployment pool, the frozen model to evaluate, and a fixed annotation budget $Y \ll X$. **(A) Headline Estimation:** a uniform-random cohort, Track A, gives an unbiased headline error $\bar{e}_A \pm \mathrm{CI}$, with silent failures entering at their true rate. [*(S1) in the protocol*] **(B) Region Grouping:** a confidence-blind error predictor scores the pool in embedding space and groups high-error inputs into candidate regions. [*(S3), (S4), (S5) in the protocol*] **(C) Error Confirmation:** disjoint labels inside each region are labeled as Track B, with one-sided tests to determine if the region's error exceeds $\bar{e}_A$ under a Benjamini–Hochberg correction [*(S6), (S7) in the protocol*]. Outputs (right): the unbiased headline with its interval, a confidence-trustworthiness verdict, an embedding map of confirmed failures, the failure modes (location × silent/loud), and the annotated, reusable cohort. The embedding space is high-dimensional; 2D projection shown for illustration.

Headline Estimator. In (S1), the headline is the mean error over the uniform cohort, unbiased for $\bar{e}$ because the cohort is not selected on any model output.

Region Grouping. Discovery (S4) trains a confidence-blind error predictor (a linear logistic probe) on the Track A embeddings computed in (S3); grouping (S5) scores the whole pool and clusters high-scoring inputs into candidate regions with a count-free method (default: GMM–BIC). We use the model's own penultimate features as the representation; a generic backbone (DINOv2/DINOv3 [22,29]) is a substitute when the model's features are unavailable.

Error Confirmation. In (S7-S8) Track B confirms each region on labels disjoint from those used to select it, which is what removes the circularity of same-

Table 1. Test bed: five model and dataset pairs. Correctness is the per-input score the auditor reveals when spending one label: a 0/1 indicator for the binary tasks (exact match for VQA), a continuous per-frame mIoU for segmentation. N is the pool size.

Pair	Model	Dataset	Performance	Correctness
Phase (CAMMA)	CAMMA baseline	Cholec80	acc. 0.55	binary, $N \approx 1200$
Phase (DINOv2-MLP)	DINOv2 + MLP	Cholec80	acc. 0.61	binary, $N \approx 1200$
Phase (MS-TCN)	DINOv2 + MS-TCN	Cholec80	acc. 0.79	binary, $N \approx 1200$
VQA	PitVQA-Net	PitVQA	acc. 0.73	binary, $N \approx 1000$
Segmentation	FCN-ResNet50	CholecSeg8k	mIoU 0.68	continuous, $N \approx 230$

cohort testing. Multiplicity across the m regions is controlled by Benjamini–Hochberg [5] at false-discovery level q. The silent fraction of a confirmed region is the share of its failures with model confidence at or above a threshold τ (default $\tau = 0.90$, confidence being the model's largest softmax probability).

Budget sizing and intervals. Given a total budget of Y labels split by a fraction α to Track A, the headline uses $n_A = \alpha Y$ labels and the remaining $Y_B = (1-\alpha)Y$ confirm regions. Each region is confirmed with $n_{\min}$ fresh labels, so the audit can afford $m = Y_B / n_{\min}$ regions. We set $n_{\min} = 25$, which holds a region's standard error to at most 0.10. The budget is sized headline-first: n_A is set by the precision a site requires of the headline, and every label beyond that buys discovery, one confirmable region per $n_{\min}$ labels. Experiments use $Y = 0.20\,N$ and $\alpha = 0.40$, which leads to headline standard errors of 3–5 pp and at least one confirmable region on every pair. Headline and region intervals use the Wilson score interval for the binary tasks and a bootstrap for mIoU.

Datasets and simulation. We use five model and dataset pairs (Table 1): three phase-recognition models on Cholec80 [31] (a CAMMA baseline at 0.55, DINOv2 [22] features with an MLP head at 0.61, and DINOv2 features with an MS-TCN [2] temporal head at 0.79), a VQA model (PitVQA-Net on PitVQA [11]) scored by exact match at 0.73, and a segmentation model (FCN-ResNet50 on CholecSeg8k [12]) scored by per-frame mIoU at 0.68. Each is fully labeled, giving the true rate against which estimates are measured; the audit is simulated by an oracle that returns an input's correctness only when a unit of budget is spent (so every reported quantity is what a real auditor would obtain from the same number of annotations). Standard errors, coverages, and false-confirmation rates are computed over 300 independent simulated audits.

3 Results

We apply the audit protocol to all five model and dataset pairs of Table 1. Because every pool is fully labeled, we know each site's true error rate exactly, so every number below is measured against that ground truth and averaged over

hundreds of independent simulated audits, each spending 20% of the dataset size in budget. The results are organized around the three properties that make this problem hard (rather than around the protocol's steps): that a uniform cohort beats confidence-weighted sampling for the headline, that silent failures are large and reachable only without confidence, and that a discovered region is real only once confirmed on fresh labels. A final paragraph reports the deliverable the audit returns at a fixed budget.

A uniform cohort gives a trustworthy headline; confidence-weighted sampling does not. Does spending labels on the frames a model is unsure about buy a better headline than a plain uniform sample? In Table 2, we compare three couples of sampling design and estimator for the site's error rate at the same budget: uniform sampling with the plain sample mean (ours); confidence-based sampling, which labels the inputs the model is least sure about, corrected by inverse-probability weighting (naive IPW); and the same confidence-based sampling corrected by a prediction-powered estimator that also trims the largest weights (Active PPI). The comparison is not exhaustive over couples (uniform sampling could equally be paired with a prediction-powered estimator), but it contrasts the two strategies open to an audit: either draw samples to label uniformly, or steer the labels by the model's confidence toward suspected failures. The uniform couple is unbiased by construction (within 0.6 pp), moves little from one audit to the next (1.7–2.8 pp), and its 95% interval contains the true rate in 96 to 99 of every 100 audits on all five pairs. Both confidence-sampled couples are also unbiased on average, yet both fail on the single audit that a real site actually runs. On the two overconfident models especially, DINOv2-MLP phase and VQA, coverage falls to 0.68–0.70 and the worst single-audit error reaches 37–61 pp, because a confident input sampled against long odds enters the weighted average with a huge corrective weight. They break on exactly the models whose confidence cannot be trusted, which are the models an audit most needs to catch; the uniform cohort is immune because it reads no confidence at all. The instability traces to the confidence-based sampling design rather than to the estimators: swapping IPW for the prediction-powered correction leaves coverage and worst-case error nearly unchanged (Table 2). The uniform cohort is also cheap: its standard error drops below 5 pp after 50–100 labels and below 3 pp after 180–240 on the classification tasks. This same cohort then does three jobs, supplying the headline, training the failure predictor, and setting the background rate for confirmation.

Silent failures are large, and only a confidence-blind signal reaches them. The error predictor of (S4) is a plain classifier trained on the Track A frames and their labels, that scores the rest of the pool by resemblance in embedding space to the frames that failed. We steer by this signal rather than by the models' confidence, as the weighted estimators do, because the failure map must capture every kind of failure, loud and silent alike. Our experiments in this paper hold the full ground truth, so we can measure how much is at stake in that choice.

Table 2. Headline estimators at the deployed budget (20% of the pool), per model–dataset pair, over 300 simulated audits. Coverage is the fraction of audits whose 95% interval contained the true rate; worst is the largest absolute single-audit error (pp). All three estimators are unbiased ($|\text{bias}| \leq 0.6$ pp); the uniform single-audit spread is 1.7–2.8 pp throughout, versus 4–16 pp for the two confidence-weighted estimators. Both confidence-weighted estimators lose coverage and produce extreme single-audit errors on the two overconfident pairs, DINOv2-MLP and VQA; the uniform cohort ignores confidence and stays reliable on all five. At the smallest budget tested, the worst naive-IPW audit reached 171 pp.

	Uniform (ours)		**Naive IPW**		**Active PPI**	
Pair	Cov.	Worst	Cov.	Worst	Cov.	Worst
Phase (CAMMA)	0.99	7	0.95	15	0.95	15
Phase (DINOv2-MLP)	0.98	7	0.68	65	0.68	61
Phase (MS-TCN)	0.96	7	0.83	26	0.82	26
VQA	0.98	9	0.70	37	0.70	37
Segmentation	0.97	6	0.94	15	0.94	9

Taking the least-uncertain third of frames as "confident", we find that confident-wrong predictions are 12–20% of all errors, a share that does not shrink as the model improves and peaks on our strongest phase model. Against the model's own softmax at a 0.90 threshold it is larger still, 22–61%, itself a measure of how overconfident these models are. Because these frames carry the same high confidence as correct predictions, a top-confidence ranker captures none of them, and uncertainty-weighted samplers under-sample these failures. Nothing in (S4) training targets silent failures specifically, but we find that its score separates wrong from right among the confident frames at an AUC of 0.66–0.75 (0.77 on segmentation with M's own features): enough to enrich a region, not to flag an individual prediction. End to end, confirmed regions recover up to 25% of silent failures. Finally, we find that silent and loud failures are not separate islands in the embedding: regions seeded from loud failures and regions seeded from silent ones largely coincide (Jaccard 0.41–0.75). A confirmed region is therefore a mixed pocket enriched in both kinds of failure, which is why the audit reports a silent fraction per region rather than a binary tag.

A discovered region is real only after confirmation on fresh labels. Are candidate regions, with higher error rate (formed in step (S5)) "real"? We find a candidate region by looking for a pocket of the data where the frames we happened to label failed unusually often. If we then measure that region's error rate on those same labels, we are judging it on the very evidence that made us flag it. In table 3, we show what happens when there is genuinely nothing to find: testing a region on the very labels that flagged it is circular, and confirms a spurious mode in almost every audit (from 0.81 on VQA to 1.00 on the phase models). Two intermediary steps fix it. Drawing a disjoint handful of labels inside each region (Track B) breaks the circularity and by itself brings the false-confirmation rate down to

Table 3. Confirmation validity, per model–dataset pair, at the operating budget over 200 null replicates (errors permuted, so every region's true rate equals the site average and any confirmation is spurious). "Naive" reuses the discovery labels; "Track B only" draws fresh disjoint labels with no correction; "Track B + BH" adds the Benjamini–Hochberg correction at $q = 0.05$, the fraction of confirmations we accept as false. Fresh labels remove the selection bias; the correction removes the multiplicity inflation. For segmentation the budget affords only $m = 1$ confirmable region, so BH has no multiplicity to correct; its 0.14 rate reflects the one-sided bootstrap test for mIoU running anticonservative at $n_{\min} = 25$ labels, a small-sample limitation of the bootstrap calibration.

Pair	**Naive**	**Track B only**	**Track B + BH**
Phase (CAMMA)	1.00	0.10	0.07
Phase (DINOv2-MLP)	1.00	0.08	0.06
Phase (MS-TCN)	1.00	0.07	0.05
VQA	0.81	0.21	0.10
Segmentation	0.94	0.14	0.14

0.07–0.21; adding the standard correction for having examined several regions at once (Benjamini–Hochberg) removes the remaining multiplicity inflation and holds the rate at 0.05–0.14: at the nominal 0.05 on the phase tasks, above it on VQA (0.10) and segmentation (0.14; see the caption of Table 3).

The audit returns a relocatable, categorised failure map at a fixed budget. The final output of the protocol is the failure map whose cost is determined before any label is spent. The deployed audit spends only the Track A share on the headline: at 20% of the pool with $\alpha = 0.40$, Track A takes $n_A = 96/80/18$ labels (phase/VQA/segmentation) for headline standard errors of 4.0–4.9, 4.8, and 2.9 pp, leaving Y_B for $m = 5/4/1$ confirmed regions. Over 300 audits, the headline interval covers the true rate at 0.95–0.97 on the binary pairs and 0.92 on segmentation. The map is relocatable: the audit returns each region's embedding coordinates and the trained predictor.

We find that the confirmed regions are recognizable failure categories. On phase they are adjacent-phase confusions, with the silent stratum in the late cleanup phases ("Gallbladder Packaging" and "Gallbladder Retraction" read as "Cleaning And Coagulation", silent fraction 0.48/0.47) while the opening phases fail loudly. On VQA they track question type: spatial, counting, and instrument questions fail loudly (0% silent), while surgical-step questions carry the silent failures. On segmentation the worst-mIoU frames concentrate on the gallbladder and hepatocystic-triangle (critical-view) structures. For VQA and segmentation, a region's location in the embedding identifies its category, since question types and anatomical structures occupy distinct pockets; for phase, location is less informative, and the category is the confusion pair. The deliverable is the headline with its interval, the confirmed modes with their category and silent fraction, the reusable cohort $A \cup B$, and the confidence verdict, itself a finding when confidence does not track error.

4 Discussion

We simulated audits of five surgical-AI models under a fixed budget. A uniform-random cohort gave a stable headline, with a single-audit spread of 1.7–2.8 percentage points and near-nominal coverage, while confidence-sampled estimators, though unbiased in the mean, swung on single audits and lost coverage precisely on the miscalibrated models an auditor most needs to check. Confident-wrong failures were 12–20% of the error mass, did not shrink as the model improved, and were invisible to any confidence-based selector, so reaching them required a confidence-blind signal. Confirming discovered regions on disjoint labels under a multiplicity correction kept false confirmations near their target. The main limitation is recall: at a 20% budget the audit confirms only up to 25% of silent failures, so the failure map is a lower bound, not a complete inventory, and never a license to trust an individual confident prediction. Because the audits here are simulated on public benchmarks, they establish the protocol itself rather than the site-specific drift it is built to measure. Raising the recall fraction, through stronger error representations or budget-adaptive region sizing, is the natural next step, alongside evaluation on true multi-site deployment data and the comparison of alternative discovery signals, such as slice-discovery methods, within the same audit.

Our approach aligns with evolving post-deployment oversight: under the FDA's final Predetermined Change Control Plan guidance [32], a headline with explicit acceptance criteria can serve a Modification Protocol's performance evaluation, and the confirmed modes supply the bias and rollback evidence an Impact Assessment needs, a fit that extends to EU Medical Device Regulation post-market surveillance and the EU AI Act's post-market monitoring of high-risk systems [1]. Run at intervals it supports continuous surveillance at bounded cost; run across sites it yields comparable headlines and maps.

Ultimately, our method turns a fixed annotation budget into a single audit that hands a site both a defensible performance number and a statistically valid map of where and how a model fails, at a label cost it can afford.

Disclosure of Interests. Y. Sakref and L. M. Knowlton are inventors on a provisional patent application related to the method described in this work, filed through Stanford University's Office of Technology Licensing. C. Nobuhara is a scientific advisor for Aimbient, Inc. J. Jopling is a senior officer for Apprentis, Inc.

Code availability. The audit protocol is fully specified in Section 2; code is available from the corresponding author upon reasonable request.

References

1. Regulation (EU) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) (Jun 2024)
2. Abu Farha, Y., Gall, J.: MS-TCN: Multi-stage temporal convolutional network for action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3575–3584 (2019)
3. Angelopoulos, A.N., et al.: Prediction-powered inference. Science **382**(6671), 669–674 (Nov 2023)
4. Ash, J.T., et al.: Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds (Feb 2020), arXiv:1906.03671 [cs.LG]
5. Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) **57**(1), 289–300 (Jan 1995)
6. Bolocan, V.O., et al.: Artificial Intelligence in Prostate MRI: Current Evidence and Clinical Translation Challenges-A Narrative Review. Journal of Imaging **11**(10), 335 (Sep 2025)
7. d'Eon, G., et al.: The Spotlight: A General Method for Discovering Systematic Errors in Deep Learning Models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 1962–1981. FAccT '22, Association for Computing Machinery, New York, NY, USA (Jun 2022)
8. Eyuboglu, S., et al.: Domino: Discovering Systematic Errors with Cross-Modal Embeddings (May 2022), arXiv:2203.14960 [cs.LG]
9. Gong, J., et al.: Using deep learning to identify the recurrent laryngeal nerve during thyroidectomy. Scientific Reports **11**(1), 14306 (Jul 2021)
10. Guo, C., et al.: On Calibration of Modern Neural Networks. In: Proceedings of the 34th International Conference on Machine Learning. pp. 1321–1330. PMLR (Jul 2017)
11. He, R., et al.: PitVQA: Image-Grounded Text Embedding LLM for Visual Question Answering in Pituitary Surgery. In: Linguraru, M.G., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 488–498. Springer Nature Switzerland, Cham (2024)
12. Hong, W.Y., et al.: CholecSeg8k: A Semantic Segmentation Dataset for Laparoscopic Cholecystectomy Based on Cholec80 (Dec 2020), arXiv:2012.12453 [cs.CV]
13. Kore, A., et al.: Empirical data drift detection experiments on real-world medical imaging data. Nature Communications **15**(1), 1887 (Feb 2024)
14. Laplante, S., et al.: Validation of an artificial intelligence platform for the guidance of safe laparoscopic cholecystectomy. Surgical Endoscopy **37**(3), 2260–2268 (Mar 2023)
15. Lasko, T.A., et al.: Why do probabilistic clinical models fail to transport between sites. npj Digital Medicine **7**(1), 53 (Mar 2024)
16. Li, P., et al.: Robust Sampling for Active Statistical Inference (Oct 2025)
17. Lin, J.C., et al.: Benefit-Risk Reporting for FDA-Cleared Artificial IntelligenceEnabled Medical Devices. JAMA Health Forum **6**(9), e253351 (Sep 2025)
18. Madani, A., et al.: Artificial intelligence for intraoperative guidance: Using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. Annals of Surgery **276**(2), 363–369 (2022)
19. Maier-Hein, L., et al.: Surgical data science – from concepts toward clinical translation. Medical Image Analysis **76**, 102306 (Feb 2022)

20. Mascagni, P., et al.: Artificial Intelligence for Surgical Safety: Automatic Assessment of the Critical View of Safety in Laparoscopic Cholecystectomy Using Deep Learning. Annals of Surgery **275**(5), 955–961 (May 2022)
21. Mascagni, P., et al.: Computer vision in surgery: from potential to clinical value. npj Digital Medicine **5**(1), 163 (Oct 2022)
22. Oquab, M., et al.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (Jul 2023)
23. Protserov, S., et al.: Development, deployment and scaling of operating room-ready artificial intelligence for real-time surgical decision support. npj Digital Medicine **7**(1), 231 (Sep 2024)
24. Roschewitz, M., et al.: Automatic correction of performance drift under acquisition shift in medical image classification. Nature Communications **14**(1), 6608 (Oct 2023)
25. Roschewitz, M., et al.: Automatic Dataset Shift Identification to Support Safe Deployment of Medical Imaging AI. In: Gee, J.C., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. pp. 67–76. Springer Nature Switzerland, Cham (2026)
26. Sakref, Y., et al.: Empowering surgeons with integrated synthetic data: solutions for mastering complex clinical scenarios. npj Digital Medicine **9**(1), 513 (Apr 2026)
27. Sener, O., Savarese, S.: Active Learning for Convolutional Neural Networks: A Core-Set Approach (Jun 2018), arXiv:1708.00489 [stat.ML]
28. Settles, B.: Active Learning Literature Survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences (2009)
29. Siméoni, O., et al.: DINOv3 (Aug 2025), arXiv:2508.10104 [cs.CV]
30. Sun, Y., et al.: Out-of-Distribution Detection with Deep Nearest Neighbors. In: Proceedings of the 39th International Conference on Machine Learning. pp. 20827–20840. PMLR (Jun 2022)
31. Twinanda, A.P., et al.: EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. IEEE Transactions on Medical Imaging **36**(1), 86–97 (Jan 2017)
32. U.S. Food and Drug Administration: Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions (Aug 2025)
33. Varghese, C., et al.: Artificial intelligence in surgery. Nature Medicine **30**(5), 1257–1268 (May 2024)
34. Wagner, M., et al.: Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark. Medical Image Analysis **86**, 102770 (May 2023)
35. Wu, E., et al.: How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nature Medicine **27**(4), 582–584 (Apr 2021)
36. Zrnic, T., Candes, E.: Active Statistical Inference. In: Proceedings of the 41st International Conference on Machine Learning. pp. 62993–63010. PMLR (Jul 2024)