

Reproducibility-Aware Evaluation of Hierarchical Multiple- Instance Learning for Patient-Level Thrombosis Detection from Compression Ultrasound

Lei Jiao

Xuzhou Mining Group General Hospital

Zhen Zhou

Xuzhou Mining Group General Hospital

Hongsheng Yin

China University of Mining and Technology

Jinhu Ma

Xuzhou Mining Group General Hospital

Chuanhe Qian

Xuzhou Mining Group General Hospital

Tao Li

Xuzhou Mining Group General Hospital

Hao Zhang

7486516@qq.com

Xuzhou Stomatological Hospital

Research Article

Keywords: compression ultrasound, multiple-instance learning, hierarchical aggregation, patient-level detection, reproducibility evaluation, internal cross-validation

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10881238/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

To provide a reproducibility-aware evaluation of multi-site cine multiple-instance learning (MS-CineMIL) versus a fixed 8-frame imaging baseline under patient-level five-fold cross-validation, and to examine whether the nominal ROC AUC difference reproduces after field completion and runtime environment control. The study delineates aggregation applicability boundaries rather than presupposing a stable advantage.

Methods

The public ThrombUS+ Dataset #1 Training Set (v1.0; 594 subjects, 2,341 DICOM cine clips) was used. Thrombosis is a per-site label, so the patient-level endpoint was defined as engineering-positive: raw Thrombosis = 1 for any record of the same SubjectID (81/594). After quality control, 2,327 cines were retained. Partitions used SubjectID isolation. Primary metric: ROC AUC; secondary: PR AUC, Brier score. Model differences were tested by 10,000-resample patient-level paired bootstrap (empirical two-sided P values; Bonferroni $\alpha = 0.0167$).

Results

In the V1 historical frozen analysis, the ROC AUC difference of MS-CineMIL versus the fixed 8-frame baseline was + 0.0740 (95% CI 0.0043 to 0.1454, $P = 0.039$), failing Bonferroni correction. In the same-environment controlled comparison, the difference was - 0.0283 (95% CI - 0.1177 to 0.0608, $P = 0.538$) with a larger Brier error (+ 0.0374, $P < 0.001$). The baseline re-run yielded 0.6390; the complete model ROC AUC was 0.6419. The between-run baseline shift (0.0711) matched the nominal gain (+ 0.0740); no stable advantage emerged.

Conclusions

The V1 frozen difference did not pass multiple-comparison correction, was not reproduced under controlled comparison, and was accompanied by a larger Brier error. Evidence supports boundary auditing of model components, field completeness, and aggregation strategies, but not stable superiority, clinical deployment, external applicability, or novelty claims.

1 Introduction

Lower-extremity venous thrombosis can lead to adverse outcomes including pulmonary embolism, recurrent venous thromboembolism, and post-thrombotic syndrome [1, 2]. Compression ultrasound is a key imaging modality for clinical assessment of suspected deep vein thrombosis [3–5]; its core

information resides not only in any single static image but also in the temporal changes of venous lumen morphology during probe compression. In recent years, artificial intelligence has been applied to ultrasound image analysis, operation guidance, and thrombosis-related computer-assisted detection [6–10]; however, reducing dynamic sequences to single frames or simply averaging probabilities over a few frames may omit temporal information and complementary signals across anatomical sites.

Real-world examinations typically contain dynamic sequences from multiple anatomical sites, both lateralities, and varying numbers per patient. If videos or frames are treated as independent samples for random partitioning, information from the same patient may span both training and test sets, causing performance overestimation; conversely, ignoring site and laterality may discard clinically meaningful structural information. Multiple-instance learning [11, 12] allows a patient to be treated as a "bag" and multiple cines or frames as "instances", enabling hierarchical aggregation under patient-level weak labels only. Gated attention mechanisms [13] can further learn non-uniform weights without requiring frame-level lesion annotation, although attention weights cannot be directly equated with lesion localization or causal interpretation. Subsequent work has extended MIL aggregation design in medical imaging, including transformer-based instance-correlation modeling for whole-slide pathology [14], hierarchical aggregation for microscopy classification [15], and multi-resolution hierarchical attention [16]; the component audited in this study is precisely this aggregation layer.

This study constructed MS-CineMIL (multi-site cine multiple-instance learning) based on the public ThrombUS+ training set [17]. The model first performs temporal aggregation over 8 uniformly sampled frames within each cine, then applies attention-based aggregation across multiple cines at the patient level to output a single patient-level probability. Anatomical site, laterality, and missing-site information are part of the model design; however, the original V1 ablation manifest did not carry the site and laterality fields, so the previous "without site/laterality encoding" difference cannot be used to demonstrate the true contribution of anatomical encoding. The research focus is not on demonstrating clinical deployment readiness. Instead, this study answers one question: under strict patient-level leakage-free internal five-fold validation, can hierarchical multi-frame, multi-site modeling, relative to fixed 8-frame simple aggregation, reproduce stably after field completion and runtime environment control? To this end, this paper adopts a reproducibility-aware evaluation design: V1 is treated as a historical frozen analysis, the field-completed re-runs and the same-environment controlled paired comparison are treated as independent evidence layers, and sensitivity analyses including label aggregation are reported.

2 Materials and Methods

2.1 Study Design and Data Source

This study is a retrospective secondary analysis and model development study based on public de-identified data. The data come from the ThrombUS+ Ultrasound Dataset #1 Training Set (v1.0; DOI: 10.5281/zenodo.17659415; CC BY 4.0) [17]. This analysis uses only the training set for patient-level

internal five-fold cross-validation and does not include the official hidden test set, independent external cohorts, or prospective clinical validation. The acquisition institutions, original ethics approvals, and informed consent information of the public data are documented in the dataset release notes; this study was reviewed and approved by the Biomedical Research Ethics Committee of Xuzhou Mining Group General Hospital (approval number: [2026] 080701; August 7, 2026).

2.2 Data Hierarchy and Engineering Endpoint Definition

The data hierarchy comprises subjects, cines, and frames. SubjectID serves as the unique patient-level identifier; each patient contains one or more dynamic ultrasound cines, and each cine contains multiple DICOM frames. The original audit included 594 subjects and 2,341 readable DICOM cines, with per-clip frame counts ranging from 15 to 372 (median 142). The four anatomical sites—CFV (common femoral vein), FV (femoral vein), PV (popliteal vein), and GS (great saphenous)—had 588, 589, 584, and 580 records, respectively [17].

Among them, a total of 162 video/site-level records with raw Thrombosis = 1 were distributed across the four anatomical sites CFV, FV, PV, and GS.

2.3 Technical Quality Control and Risk Closed-Loop

All 2,341 DICOM clips were first audited for readability, frame count, near-black pixel ratio, grayscale variance, and potential burned-in information; all DICOM headers were readable. Two associate chief physicians of ultrasound with 16 years of relevant experience performed technical quality review of prespecified sampled images, evaluating effective ultrasound content, black borders or empty regions, identity information, diagnostic text, measurement values, and other burned-in markers. Both reviewers were aware of the label semantics during the initial review; therefore, this process is not characterized as label-blinded. The technical review was used solely to formulate label-agnostic image retention, quarantine, and exclusion rules; it was not used to modify the raw Thrombosis field, nor does it constitute disease outcome relabeling or clinical gold-standard adjudication.

All original review records were retained. For one reviewer's mis-entered Yes/No direction on "usable ultrasound content", a derived correction with an audit trail was established while preserving the original files. Identity, diagnostic text, measurement information, and other potential shortcut risks were handled conservatively according to the union reported by the two reviewers. Ultimately, 7 expert-review-uncertain cines were quarantined; another 7 geometry-pending cines were excluded in full under the fail-closed principle, because no validated fixed occlusion or cropping scheme had been established. A total of 14 cines were excluded, and 2,327 cines were retained for formal image analysis; all 594 subjects retained at least one usable cine, and technical quality control did not modify any disease labels. The third-reviewer technical QC deliverable package comprises 50 frame-by-frame images R001–R050, 5 overview images, and a 50-row review worksheet, completed independently by chief physician Tao Li. This review was performed at the image-row level, tallying 37 retained, 8 retained-after-cropping, and 5 excluded frames; among them, R015 was originally recorded as requiring expert review and was included in the excluded frames upon independent expert confirmation; there are currently no pending expert-

review items. Identity information and diagnostic text were present in 0 images. The third-reviewer QC was used for technical image quality auditing and not for disease label relabeling, gold-standard confirmation, or disease outcome adjudication.

2.4 Image Preprocessing and Frame Sampling

Image processing followed label-agnostic and cross-fold-consistent principles. Each retained cine was deterministically uniformly sampled at 8 frames over the full timeline; sampling did not depend on labels or manual "best-frame" selection. Frames were converted to three-channel images and resized to 160×160 pixels by bilinear interpolation. The near-black pixel ratio cap was 0.99, and the minimum grayscale standard deviation was 5.0. Preprocessing, QC rules, frame cache provenance, patient fold splits, and run configurations were all hashed into run signatures. 224×224 was used only as a prespecified resolution sensitivity analysis, reusing the same patient folds, labels, and model configuration, without additional tuning, threshold optimization, or fusion.

2.5 Patient-Level Five-Fold Cross-Validation

All data partitions were performed at the SubjectID level [18]. The 594 subjects were divided into 5 folds; outer test-set sizes were 120, 119, 119, 118, and 118, with positive counts of 17, 16, 16, 16, and 16, respectively. When fold k served as the test set, the next fold served as the validation set, and the remaining three folds were used for training. The validation set was used for early stopping and best-checkpoint selection; the outer test set did not participate in model selection. The five test sets were mutually exclusive, and their union exactly covered all 594 subjects; each patient appears exactly once in the pooled out-of-fold predictions. All compared models used the same patient sets, the same labels, and the same fold numbering.

2.6 Compared Models

The structured logistic baseline used five variables—age, sex, height, weight, and thigh circumference—and explicitly excluded Compressibility to avoid using a variable near the target definition as predictive information. The fixed 8-frame imaging baseline uniformly sampled 8 frames per cine, used a MobileNetV3-Small [19] shared encoder without pretrained weights to produce cine probabilities, and then averaged the cine probabilities of all cines of the same patient to obtain patient-level predictions.

2.7 MS-CineMIL Model

MS-CineMIL treats each patient's multiple cines as a hierarchical instance bag. The 8 frames of each cine pass through a shared MobileNetV3-Small encoder to produce 576-dimensional frame features; gated temporal attention aggregates frame features into a cine representation. An 8-dimensional anatomical site embedding and a 4-dimensional laterality embedding are then added, followed by a linear layer mapping to 256-dimensional cine features. Patient-level gated attention performs non-uniform aggregation over a variable number of cines, concatenates the missingness indicator vector of the four sites CFV, FV, PV, and GS to the patient representation, and finally outputs a patient-level logit.

The temporal attention dimension and the patient attention dimension are both 128. Padding frames and padding cines are explicitly excluded via masks.

The hierarchical data flow of the model and the two attention aggregation layers are shown in Fig. 1.

2.8 Training and Model Selection

MS-CineMIL used the patient as the training unit with BCEWithLogitsLoss and a positive-class weight; the fixed 8-frame baseline used the cine as the training unit, but each cine's weight was corrected by the reciprocal of the number of cines of the same patient, preventing patients with many cines from gaining disproportionate total weight within an epoch. The imaging models used the AdamW optimizer with a learning rate of 3×10^{-4} , weight decay of 1×10^{-4} , a maximum of 30 epochs, early-stopping patience of 7, random seed 2026, and automatic mixed precision. The MS-CineMIL batch size was 2 patients. At the end of each epoch, patient-level probabilities were formed on the validation set, and the best model was saved at the epoch with the lowest patient-level binary cross-entropy. The prespecified compared models and their ablation and re-run variants were consistent in optimizer, learning rate, weight decay, maximum epochs, early-stopping patience, random seed, and automatic mixed precision settings (batch size units differ: cine for the fixed 8-frame baseline, patient for MS-CineMIL, see above), without adjusting the training budget across models; deterministic algorithm switches were not enabled in any run.

2.9 Evaluation Metrics and Statistical Analysis

The primary evaluation metric was patient-level ROC AUC; key secondary metrics were PR AUC and the Brier score [20]. The distributional characteristics of continuous variables can be assessed descriptively with the Shapiro–Wilk test, but the main model comparisons in this study are based on patient-level out-of-fold predictions and paired bootstrap [21] and do not rely on normality assumptions. All model point estimates are based on the pooled 594 out-of-fold predictions. Differences among MS-CineMIL, the fixed 8-frame baseline, and structured logistic regression were computed with a 10,000-resample patient-level paired nonparametric bootstrap using percentile 95% CIs; differences are uniformly defined as comparison model minus reference model, with negative Brier differences indicating better performance. Inferential endpoints were stratified in advance: patient-level ROC AUC is the sole primary inferential endpoint; the ROC AUC comparisons of the three prespecified pairwise model comparisons (MS-CineMIL vs fixed 8-frame baseline, MS-CineMIL vs structured logistic regression, fixed 8-frame baseline vs structured logistic regression) constitute the sole multiple-comparison correction family, with a Bonferroni-corrected two-sided significance threshold of $\alpha = 0.0167$ ($0.05/3$). PR AUC and Brier score are key secondary/supportive metrics, reported only as effect sizes, nominal 95% CIs, and empirical two-sided P values, not entering the multiple-comparison family and not stated as statistically significant after multiplicity correction. The bootstrap intervals reported in this paper are nominal 95% CIs; two-sided P values were computed directly from the same 10,000-resample bootstrap difference distribution, defined as $2 \times \min(P(\Delta \leq 0), P(\Delta \geq 0))$; this P value is an empirical two-sided P value under 10,000 resamples, with a resolution floor of approximately 0.0002. It should be noted that the empirical P value based on the bootstrap difference distribution is an exploratory inference method, and its results

may differ from those of the DeLong test or permutation tests; this paper positions it as supportive evidence and not as the sole basis for confirmatory hypothesis testing. The above confidence intervals and P values reflect only the conditional uncertainty on frozen out-of-fold predictions and do not include the uncertainty introduced by re-partitioning and retraining. Multiple-comparison interpretation follows the Bonferroni threshold above: only when a P value within the correction family falls below 0.0167 is it stated as statistically significant after multiple-comparison correction; all other cases are reported as nominal difference signals. All inferential tests were two-sided.

2.10 Ablation, Calibration, and Resolution Sensitivity Analyses

Both core ablations kept the patient folds and training protocol unchanged: one removes the anatomical site and laterality encoding; the other replaces patient-level gated attention with uniform mean pooling over all valid cines. The calibration analysis was based on the frozen MS-CineMIL out-of-fold predictions: fold-specific Platt transformations [22] were fitted on each fold's corresponding validation set and applied to the outer test fold, with both raw and calibrated Brier, expected calibration error (ECE) [23], calibration intercept, and slope reported. This procedure reuses the validation folds used for model selection and is therefore only exploratory internal calibration, not strictly nested cross-calibration. The resolution sensitivity analysis changed the input to 224×224 while keeping other key configurations consistent with the V1 historical frozen 160×160 configuration. The label aggregation sensitivity analysis, without changing folds, QC rules, or the training protocol, redefined the patient-level engineering-positive label from "any cine positive" to "majority cines positive" (label flips in 53 patients; positive count reduced from 81 to 28), retrained the complete MS-CineMIL, and reported five-fold OOF metrics, to assess the sensitivity of conclusions to the patient-level endpoint definition.

During the preparation of this work, the authors used an AI-based language model to assist with English-language drafting and editing of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

3 Results

3.1 Cohort, Technical Quality Control, and Out-of-Fold Prediction Integrity

Among the 594 subjects, 81 were engineering-positive and 513 engineering-negative. Of the 2,341 raw cines, positive records at PV, FV, CFV, and GS numbered 55, 42, 39, and 26, respectively. After technical QC excluded 14 cines in full, the input to the prespecified compared models was 2,327 cines, and all 594 subjects were retained. Among the 14 excluded cines, 5 had raw Thrombosis = 1 and 9 had Thrombosis = 0, involving 13 independent subjects (affected subjects per outer fold: 1, 2, 3, 4, and 3); no engineering-positive patient had all of their raw positive cines excluded by technical QC—that is, no "engineering-positive label retained while no raw positive cine remains in the model input" label–input mismatch

occurred (Supplementary Table S1). The five test folds were mutually independent, and their union exactly matched the patient roster; all three prespecified compared models obtained 594 unique patient-level out-of-fold predictions, with identical patient sets, labels, and folds across models. The third-reviewer deliverable package files were complete, with the 50 R-numbers corresponding one-to-one to the 50 review images; its results are reported as supplementary technical QC evidence and do not alter the existing model input boundaries. An overview of the cohort, technical quality control, and patient-level cross-validation is presented in Table 1.

Table 1
Overview of the cohort, technical quality control, and patient-level cross-validation

Item	Value
Total subjects	594
Engineering-positive / engineering-negative	81 / 513
Engineering-positive rate	13.64%
Raw DICOM cines	2,341
Cines excluded by technical QC	14
Input cines for prespecified compared models	2,327
Frames per cine	15–372; median 142
Outer test-set sizes across five folds	120, 119, 119, 118, 118
Outer test-set positive counts across five folds	17, 16, 16, 16, 16

Note: Technical quality control adjusts only the model imaging input boundaries and does not modify any raw Thrombosis field; the patient-level engineering-positive endpoint is defined by the original records and is independent of QC exclusions.

3.2 Patient-Level Performance of Prespecified Compared Models

MS-CineMIL achieved a pooled patient-level out-of-fold ROC AUC of 0.6419 (95% CI 0.5693 to 0.7104), PR AUC of 0.2261 (95% CI 0.1685 to 0.3156), and Brier score of 0.1627 (95% CI 0.1440 to 0.1829). The structured logistic regression and fixed 8-frame imaging baseline achieved ROC AUCs of 0.5928 and 0.5679, respectively. The patient-level out-of-fold performance point estimates of the three prespecified compared models are summarized in Table 2.

Table 2
Patient-level out-of-fold performance point estimates of the three
prespecified compared models (V1 historical frozen analysis)

Model	ROC AUC	PR AUC	Brier
Structured logistic regression	0.5928	0.1708	0.2389
Fixed 8-frame imaging baseline	0.5679	0.1711	0.1611
MS-CineMIL	0.6419	0.2261	0.1627

In the historical frozen V1, compared with the fixed 8-frame baseline, the ROC AUC difference of MS-CineMIL was 0.0740 (nominal 95% CI 0.0043 to 0.1454, $P = 0.039$), the PR AUC difference was 0.0550 ($P = 0.062$), and the Brier difference was 0.0016 ($P = 0.834$). This ROC AUC difference belongs to the prespecified primary inferential endpoint family and did not pass the Bonferroni-corrected threshold $\alpha = 0.0167$, constituting only a nominal signal at the historical frozen layer; the PR AUC and Brier differences are key secondary metrics, without claims of multiple-comparison-corrected significance. More importantly, in the same-environment controlled re-run after field auditing, the ROC difference of MS-CineMIL versus the fixed 8-frame re-run was -0.0283 (95% CI -0.1177 to 0.0608 , $P = 0.538$), and the Brier difference was $+0.0374$ (95% CI $+0.0228$ to $+0.0519$, $P < 0.001$; positive values indicate larger probability error). Therefore, the V1 nominal difference cannot be interpreted as a demonstrated advantage of the hierarchical model; the current interpretation is bounded by the non-reproduction in the same-environment controlled comparison (see Section 3.6). The ROC and PR difference intervals versus structured logistic regression also cross 0; its Brier comparison reflects only a difference in probability error and cannot be interpreted as validation of clinical event probabilities. The full patient-level paired-bootstrap model-difference table is provided as Supplementary Table S2 in the Supplementary Material section at the end of this manuscript. The patient-level out-of-fold ROC and PR curves of the three prespecified compared models are shown in Fig. 2 and Fig. 3, respectively.

3.3 Core Ablation Experiments

The core ablation results of the V1 historical frozen analysis are shown in Table 3 and Fig. 4: the complete model ROC AUC was 0.6419, the model without site/laterality encoding was 0.5541, and the uniform mean-pooling model was 0.5335. The "without site and laterality encoding" result reported in the original V1 cannot be interpreted as the true contribution of anatomical encoding, because the original V1 ablation manifest did not carry the site and laterality fields. This section reports only the ablation results at the V1 historical frozen layer; the independent re-runs and paired comparisons after field completion (corrected-input) are presented in Section 3.6.

Table 3
MS-CineMIL core ablation results (V1 historical frozen manifest; field-completed re-runs and their paired comparisons are presented in Section 3.6)

Model/ablation	ROC AUC	PR AUC	Brier
Complete MS-CineMIL	0.6419	0.2261	0.1627
Without site/laterality encoding	0.5541	0.1642	0.1770
Uniform mean pooling	0.5335	0.1494	0.2250

3.4 Exploratory Calibration Analysis

The frozen raw MS-CineMIL out-of-fold predictions had a Brier score of 0.1627, ECE of 0.1735, calibration intercept of -1.6008 (95% CI -1.8906 to -1.3249), and calibration slope of 0.1014 (95% CI 0.0304 to 0.2274), indicating insufficient probability calibration. After the fold-specific Platt transformation, the Brier score decreased to 0.1161 and ECE to 0.0241 ; however, ROC AUC decreased to 0.6118 , and one fold's Platt slope was negative (-0.1241). Because the calibration fitting reuses the validation folds used for model selection, and fold-specific transformations may alter the global ranking across folds, these results serve only as a risk audit and do not support using the calibrated probabilities directly for clinical threshold decisions. The full exploratory calibration results table is provided as Supplementary Table S3 in the Supplementary Material section at the end of this manuscript. The reliability curves of MS-CineMIL before and after exploratory Platt calibration are shown in Fig. 5.

3.5 Resolution Sensitivity: 160×160 versus 224×224

The 224×224 MS-CineMIL achieved a ROC AUC of 0.6065 (95% CI 0.5343 to 0.6749), PR AUC of 0.2083 (95% CI 0.1509 to 0.2862), and Brier score of 0.1873 (95% CI 0.1714 to 0.2040). Relative to the frozen 160×160 main model, the ROC AUC difference was -0.0354 (95% CI -0.1071 to 0.0345), and the PR AUC difference was -0.0178 (95% CI -0.0891 to 0.0441), neither showing a stable discrimination gain; the Brier difference was 0.0246 (95% CI 0.0085 to 0.0411), an unfavorable direction. Accordingly, 160×160 is retained as the V1 historical frozen configuration. The input resolution sensitivity analysis is summarized in Table 4, and the patient-level paired performance differences of 224×224 relative to the frozen 160×160 main model are shown in Fig. 6.

Table 4
Input resolution sensitivity analysis

Resolution	ROC AUC (95% CI)	PR AUC (95% CI)	Brier (95% CI)
160×160 (V1 historical frozen configuration)	0.6419 (0.5693 to 0.7104)	0.2261 (0.1685 to 0.3156)	0.1627 (0.1440 to 0.1829)
224×224	0.6065 (0.5343 to 0.6749)	0.2083 (0.1509 to 0.2862)	0.1873 (0.1714 to 0.2040)

3.6 Independent Re-Runs and Controlled Comparisons After Field Auditing

Because the original V1 frozen manifest lacked the Anatomical Site and Limb fields present in the raw train.csv, the research team generated a completed manifest without altering the raw data, patient-level labels, QC rules, or SubjectID folds, and used it as an independent analysis input. This re-run does not supersede Experiment V1; the results are used only to assess whether model performance and the original interpretation boundaries remain stable after true field completion. The complete model, the model without site/laterality encoding, and the uniform mean-pooling model each obtained one-per-patient OOF predictions for all 594 patients, using the same QC and cache provenance; point estimates are shown in Table 5: the complete model achieved ROC AUC 0.6107, PR AUC 0.2302, and Brier 0.1976; the model without site/laterality encoding achieved 0.5957, 0.1891, and 0.1768, respectively; the uniform mean-pooling model achieved 0.6278, 0.2516, and 0.2201, respectively. The complete model versus the model without site/laterality encoding showed a ROC difference of + 0.0150 (95% CI – 0.0633 to 0.0953), PR difference of + 0.0412 (95% CI – 0.0216 to 0.1133), and Brier difference of + 0.0208 (95% CI 0.0054 to 0.0360), with corresponding two-sided P values of 0.726, 0.206, and 0.008; the between-fold directions of ROC and PR were unstable, and the complete model's Brier was worse. These results do not support upgrading site/laterality encoding to a formal novelty mechanism.

Under the field-completed manifest, a five-fold re-run of the fixed 8-frame imaging baseline was additionally completed (this baseline consumes only the four columns File Name, SubjectID, SubjectLabel, and Fold, with input data identical to V1), yielding ROC AUC 0.6390 (95% CI 0.5702 to 0.7049), PR AUC 0.2130, and Brier score 0.1602. The difference between this re-run and the identically named V1 baseline (ROC AUC 0.5679) reached 0.0711, under identical data, random seed, and hyperparameters. Both runs were completed on the same device (both run reports show an NVIDIA GeForce RTX 4090 with 24 GB of memory, identical model and capacity), so hardware replacement can be excluded as a confounder; the difference is consistent with the software-stack difference between the two runs (V1: Python 3.11.9, PyTorch 2.7.1, CUDA 11.8; re-run: Python 3.13.14, PyTorch 2.11.0, CUDA 12.8) and the resulting GPU nondeterminism and divergent early-stopping trajectories (fold 0 early-stopping epochs were 9 and 24); however, cuDNN, driver versions, and deterministic switch settings were not fully recorded, so the drift cannot be uniquely attributed to the software stack. This difference is of the same order as the nominal gain of MS-CineMIL versus the fixed 8-frame baseline in the V1 historical frozen analysis (+ 0.0740), suggesting that the latter is sensitive to the runtime environment and should not be over-interpreted.

The per-fold report files of every formal training automatically record run signatures—including Python, PyTorch, CUDA runtime, pydicom, scikit-learn, and NumPy versions, operating system platform, GPU model and memory, input manifest hash, and run identifier. These files are archived with the code and outputs for review. cuDNN version, GPU driver version, thread counts, and deterministic-algorithm switch

status are not included in this signature (deterministic algorithms were not enabled in either run); this recording gap is disclosed as such.

Table 5
Patient-level out-of-fold performance of independent re-runs after field auditing (field-completed manifest; 594 patients, 81 positives; internal five-fold cross-validation)

Model (corrected-input)	ROC AUC	PR AUC	Brier
Complete MS-CineMIL	0.6107	0.2302	0.1976
Without site/laterality encoding	0.5957	0.1891	0.1768
Uniform mean pooling	0.6278	0.2516	0.2201
Fixed 8-frame imaging baseline (re-run)	0.6390	0.2130	0.1602

The results of the same-environment controlled paired comparison (10,000-resample patient-level paired bootstrap, seed = 20260802, methods consistent with the V1 historical frozen analysis) are as follows. The ROC AUC difference of the complete MS-CineMIL versus the fixed 8-frame re-run baseline was - 0.0283 (95% CI - 0.1177 to 0.0608, two-sided P = 0.538), the PR AUC difference was + 0.0172 (95% CI - 0.0591 to 0.0979, P = 0.664), and the Brier difference was + 0.0374 (95% CI + 0.0228 to + 0.0519, P < 0.001, positive values indicating larger probability error). The uniform mean-pooling model versus this baseline showed the same direction (ROC AUC difference - 0.0112, two-sided P = 0.778; Brier difference + 0.0599, P < 0.001). After eliminating the environmental confounder, the nominal ROC gain in the V1 historical frozen analysis did not reappear, and the Brier difference directions of both multiple-instance variants were unfavorable (Brier 0.1976 and 0.2201 vs 0.1602). Therefore, the + 0.0740 nominal ROC difference was not reproduced under controlled conditions, its magnitude is consistent with runtime environment drift, and it should not be interpreted as a robust advantage of the hierarchical structure; this non-reproduction constitutes the interpretation boundary of this paper for the V1 nominal gain.

3.7 Label Aggregation Strategy Sensitivity Analysis

After redefining the patient-level engineering-positive label from "any cine positive" to "majority cines positive", labels flipped in 53 patients and the positive count decreased from 81 to 28 (positive rate 4.7%), with folds, QC rules, and the training protocol unchanged. The five-fold OOF results of the retrained complete MS-CineMIL were: ROC AUC 0.6406 (95% CI 0.5397 to 0.7371), PR AUC 0.0994, and Brier score 0.1585. The ROC AUC point estimate is close to that of the V1 historical frozen analysis under the any-cine-positive endpoint (0.6419), indicating only that the ranking point estimate did not shift substantially under this single aggregation rule; however, the sharp decrease in positive counts substantially lowered PR AUC, and this analysis is only a sensitivity check under a single aggregation rule and does not constitute validation of label semantics correctness.

3.8 Supplementary Exploration: Temporal TCN Seed Results and Evidence Status

This section is a supplementary exploratory analysis. The TCN architecture implementation, training, and seed protocol scripts are available in the public code archive (Zenodo DOI: 10.5281/zenodo.22043843; v2.7 release) under the src and scripts directories, and the per-fold run reports of the three seeds are in the results/sensitivity directory. The patient-level five-fold OOF results of TCN [24] across three independent seeds were: for seed 2026, ROC AUC, PR AUC, and Brier of 0.5672, 0.1859, and 0.1942, respectively; for seed 2027, 0.6245, 0.2046, and 0.1987; for seed 2028, 0.5704, 0.1854, and 0.1951. The descriptive mean $\pm$ sample SD of the three-seed point estimates were ROC AUC 0.5874 ± 0.0322 , PR AUC 0.1919 ± 0.0110 , and Brier 0.1960 ± 0.0024 . This SD represents only the between-seed fluctuation of point estimates, does not substitute for patient-level confidence intervals, and does not support conclusions of stable superiority.

3.9 Third-Reviewer Technical QC Deliverable Results

The third-reviewer deliverable package comprises 1 worksheet, 50 frame-by-frame images R001–R050, and 5 overview images. The worksheet and image files correspond 50/50 one-to-one, and the third reviewer completed 50/50 of the main review rows. The third reviewer is chief physician Tao Li, who completed the worksheet independently. The final dispositions were 37 retained, 8 retained-after-cropping, and 5 excluded frames; among them, R015 was originally recorded as requiring expert review and was included in the excluded frames after independent expert confirmation, so there are currently no pending expert-review items. Identity information was present in 0 images, diagnostic text in 0, measurement values in 9, and other burned-in text/markers in 12. These results indicate that the prespecified technical image review deliverables and frame-by-frame evaluation have been completed, but they cannot be interpreted as disease label validation, three-reviewer agreement across the full cohort, or clinical validation.

4 Discussion

4.1 Principal Findings

This study completed strict patient-level five-fold internal validation on 594 subjects and 2,327 QC-passed dynamic compression ultrasound cines. In the V1 historical frozen analysis, the ROC AUC difference of MS-CineMIL versus the fixed 8-frame baseline was + 0.0740 (95% CI 0.0043 to 0.1454, empirical two-sided $P = 0.039$), but this difference did not pass Bonferroni correction ($\alpha = 0.0167$), and the PR AUC and Brier difference intervals crossed 0; more importantly, the gain did not reappear in the same-environment controlled paired comparison (Section 3.6; ROC difference – 0.0283, two-sided $P = 0.538$) and the Brier difference direction of MS-CineMIL was unfavorable (+ 0.0374, empirical two-sided $P < 0.001$; supportive evidence only, not a multiple-comparison-corrected conclusion). Between-run and cross-validation variance has been documented as a first-order concern in machine learning evaluation [25–27] and in comparing supervised classifiers [28]. The present study provides a patient-level medical-imaging example in which a nominally significant gain ($P = 0.039$) disappears under same-environment control, with the between-environment shift in the fixed 8-frame baseline ROC AUC alone (0.6390 vs

0.5679 in V1) of the same order of magnitude as the nominal model gain of + 0.0740. Sensitivity to the AUC comparison method (DeLong covariance-based vs bootstrap) [29] and to the choice of statistical test for comparing classifiers [28] was therefore examined; the bootstrap was retained as the primary tool because it yields empirical confidence intervals and two-sided P values from the same difference distribution for all three endpoints (ROC AUC, PR AUC, Brier), whereas DeLong's test is defined only for ROC AUC. Accordingly, the current evidence does not support a robustly reproducible discrimination gain from hierarchical modeling, and the main contribution of this paper is repositioned as threefold: (i) a patient-level, leakage-free evaluation and provenance-auditing framework for dynamic multi-site ultrasound; (ii) an empirical demonstration that between-run variance can reach the same order of magnitude as nominal model gains, consistent with the broader literature on machine-learning evaluation variance [25–27]; and (iii) explicit evidence boundaries that distinguish internal aggregation auditing from claims of stable superiority or clinical utility. The absolute ROC AUC of the model is 0.6419 (V1) and 0.6107 (corrected-input), still at a limited level; no internal methodological observation should be exaggerated into clinical usability.

In the indirect comparison with the published AutoDVT method, Kainz et al. [7] reported video-level ROC AUCs of approximately 0.94 (single frame) to 0.96 (sequence); however, that study differs from this one in data source, label definition, partition unit, and evaluation level, and its results were not obtained under this study's ThrombUS+ patient-level five-fold framework. Therefore, the two sets of AUCs cannot be ranked directly against each other. This study obtained MS-CineMIL ROC AUC 0.6419 under the stricter patient-level out-of-fold evaluation; this gap suggests that a substantial divide may exist between video-level random-partition results and patient-level generalization results. This gap highlights the importance of patient-level isolation and label semantics auditing. This divide is consistent with the performance overestimation from video-level random partitioning reported in the literature [18], suggesting that evaluation level and partition strictness substantially affect ultrasound AI performance estimates. This comparison serves only as methodological background and does not constitute head-to-head validation under the same data, protocol, and endpoint conditions.

4.2 Value of Dynamic Information and Patient-Level Hierarchy

The fixed 8-frame baseline already uses multi-frame input but adopts simpler probability aggregation at the cine and patient levels. MS-CineMIL establishes non-uniform aggregation at the frame, cine, and patient levels. Under the V1 historical manifest, uniform mean pooling performed worse than the complete model, but this direction reversed in the corrected-input after field auditing: the uniform mean-pooling ROC AUC of 0.6278 and PR AUC of 0.2516 exceeded the complete model's 0.6107 and 0.2302, while its Brier score was worse (0.2201 vs 0.1976). For this reason, the current results cannot demonstrate that non-uniform aggregation is stably superior to uniform mean pooling; the direction of the aggregation-strategy comparison did not remain stable across input field completeness and run configurations, and these observations should be treated as aggregation-boundary and reproducibility auditing rather than mechanistic superiority evidence. The same-environment controlled comparison

further showed that MS-CineMIL was not superior to the fixed 8-frame baseline, reinforcing that the value of non-uniform aggregation under the present evidence remains an open question.

4.3 Anatomical Site and Laterality Encoding

The original V1's "without site and laterality encoding" difference cannot demonstrate that anatomical context participates in discrimination, because the original manifest lacked the site and laterality fields. The independent re-run after field completion has been completed: the complete model versus the encoding-removed model showed a ROC difference of + 0.0150 (95% CI - 0.0633 to 0.0953), PR difference of + 0.0412 (95% CI - 0.0216 to 0.1133), and Brier difference of + 0.0208 (95% CI 0.0054 to 0.0360). Therefore, this paper does not treat this result as definitive evidence of anatomical encoding contribution. The uniform mean-pooling ablation under V1 once suggested that patient-level non-uniform aggregation might have value, but this suggestion was not reproduced in the corrected-input re-run (uniform mean-pooling ROC AUC exceeding the complete model) or the same-environment controlled comparison, and therefore does not constitute evidence supporting the value of non-uniform aggregation. Its robustness still requires testing in independent data for site distribution shifts, true missing sites, and device domain shift.

4.4 Probability Calibration and Clinical Interpretation Boundaries

The raw model's calibration slope is far below 1, indicating overly extreme predictions or cross-fold heterogeneity. The exploratory Platt transformation reduced the Brier score and ECE, but this procedure is not strictly nested cross-calibration, and one fold's mapping slope was negative, indicating unstable validation-set rankings in some folds. Thus, discrimination improvement cannot substitute for probability calibration [30]. If the model is to be used for clinical risk communication or threshold decisions in the future, the calibration mapping should be re-estimated on completely independent calibration and external test sets, with clinical thresholds locked in advance.

4.5 Resolution Sensitivity Results

Increasing the input from 160×160 to 224×224 did not yield a stable discrimination gain, and the Brier score worsened. This result does not mean that high resolution is ineffective for all ultrasound tasks; rather, it indicates that under the current sample size, label granularity, model capacity, and preprocessing conditions, additional pixel information did not translate into reproducible patient-level gains. Freezing 160×160 as the V1 historical frozen configuration both conforms to the prespecified rules and reduces computational cost and post hoc optimization risk. This paper did not systematically record training time, inference time, or peak memory per configuration, and therefore draws no quantitative conclusions based on computational efficiency metrics.

4.6 Outlook on Clinical Workflow Integration

Within strictly bounded evidence limits, hierarchical multiple-instance modeling may in the future serve as an adjunct reading direction after sonographers complete compression ultrasound examinations,

rather than as an independent diagnostic tool. MS-CineMIL has not shown a robust advantage over the simple aggregation baseline under the current evidence; this section only describes the general potential directions of this class of methods and does not claim the applicability or superiority of this model. The current internal OOF ROC AUC is 0.6419, and label semantics, probability calibration, and external applicability remain limited; therefore, clinical decision-making cannot be directly changed on this basis. Potential directions — all premised on independent external validation, rigorous calibration, pre-locked thresholds, prospective impact assessment, and clearly assigned human review responsibility — include automatic flagging of patients for focused review after examination completion, stratified screening combined with clinical pretest-probability assessment such as the Wells score [31], and cine-level information ranking in multi-site examinations to help physicians prioritize sites likely to contain more discriminative information. Model outputs should be regarded as second-reader prompts, not as replacements for sonographers or the clinical gold standard.

4.7 Strengths

The strengths of this study include: first, strict use of SubjectID for training, validation, and test partitioning, preventing frames or cines of the same patient from crossing folds; second, all models yield one-per-patient out-of-fold predictions compared with patient-level paired bootstrap; third, complete retention of original QC records, derived corrections, exclusion lists, frame cache provenance, and run signatures. Fourth, no threshold optimization on out-of-fold predictions and no selective reporting of resolutions or ablation results. Fifth, explicit distinction among technical QC, dataset labels, and the clinical gold standard, avoiding exaggerating expert technical review into disease outcome confirmation; sixth, proactive disclosure of performance drift associated with runtime environment differences, and completion of a same-environment controlled paired comparison to delineate the interpretation boundaries of the nominal gain, providing a public example of between-run variance auditing for ultrasound AI research. The paper structure and reporting boundaries follow the TRIPOD + AI and CLAIM recommendations [32, 33], and the PROBAST tool was referenced for bias risk assessment [34]. A checklist-based self-assessment against the TRIPOD + AI and CLAIM items was completed during manuscript finalization, and the completed checklists can be provided to editors and reviewers upon request.

4.8 Limitations

This study has several important limitations. First, all model results come from internal cross-validation on a single public training set, without independent external validation or official blinded testing. Second, the patient-level engineering labels were formed by aggregating the raw Thrombosis field. The data publisher stated in communications with the research team that this field is a per-video, per-anatomical-site label derived from a combination of eCRF physician reports and the corresponding ultrasound videos, and that patient-level partitioning should be performed by Patient ID (relevant correspondence available upon reasonable request). However, the publisher also explicitly stated that this preliminary dataset has not completed independent label adjudication. Consequently, the 81 engineering-positive cases cannot be equated with DVT cases confirmed by an independent endpoint committee, nor

characterized as an independently adjudicated gold standard. The dataset paper [17] uses the GS field to describe great saphenous junction–related sites; this study follows the field values in the current manifest and does not further infer patient-level clinical endpoints from them. Third, the initial two human technical reviewers were aware of the label semantics; the third-reviewer QC was completed independently by chief physician Tao Li and constitutes a prespecified technical quality audit of 50 enriched images, not a validation of disease label correctness, three-reviewer agreement across the full cohort, or gold-standard adjudication. Fourth, the R-numbers of the third-reviewer deliverable correspond one-to-one with the image files, but the images do not contain SubjectID, CineKey, or original DICOM paths; therefore, its results are not used to modify frozen model inputs or automatically trigger sensitivity recomputation. Fifth, only 81 patients are positive; model intervals remain wide, and between-fold performance fluctuates. Sixth, clinical decision thresholds were not locked in advance, so sensitivity, specificity, and rule-out capability cannot be reported or interpreted. Seventh, the exploratory Platt calibration is not a strictly nested design. Eighth, the nondeterminism of deep learning training on GPUs means that identical data, seeds, and hyperparameters can produce considerable performance drift across framework versions: the fixed 8-frame baseline re-run under the field-completed manifest (ROC AUC 0.6390) differs from the identically named V1 baseline (0.5679) by 0.0711, the same order as the nominal gain in the V1 historical frozen analysis. The + 0.0740 ROC difference should therefore be interpreted cautiously against the background of between-run variance. The gain was not reproduced in the same-environment controlled re-run and the Brier was worse (two-sided $P < 0.001$), further supporting this cautious stance. Finally, model attention weights can only describe aggregation behavior and cannot be interpreted as lesion localization or causal mechanisms.

4.9 Future Research

Future work should prioritize validating MS-CineMIL in independent data with adjudicated labels and well-defined clinical endpoint semantics, reporting external test performance under the same patient-level partitioning principles [35]. Comparisons with published methods such as AutoDVT should preferentially adopt reproduction experiments with the same data, patient-level partitioning, endpoints, and evaluation metrics; where code or weights are unavailable, only the indirect methodological comparison in this paper is retained, without numerical ranking. Patient-level labels driven by clinical endpoint definitions should be established; model comparisons should incorporate multi-seed and multi-environment repeated training, with between-run variance reported as routine context for effect size interpretation; probability calibration and pre-locked thresholds should be evaluated under complete separation of training, calibration, and testing; robustness analyses under device domain shift, missing sites, and site distribution changes should be further conducted. If real patient-level paired data of structured clinical variables and dynamic ultrasound become available, late fusion studies could be conducted without concatenating with the clinical cohort of Paper 1.

5 Conclusion

This study completed an internal five-fold evaluation of dynamic compression ultrasound under a patient-level SubjectID isolation framework. In the V1 historical frozen analysis, MS-CineMIL achieved a ROC AUC of 0.6419, with a nominal difference of + 0.0740 versus the fixed 8-frame baseline ($P = 0.039$), which did not pass Bonferroni correction; in the same-environment controlled paired comparison, the ROC difference was $- 0.0283$ ($P = 0.538$) and the Brier difference was + 0.0374 ($P < 0.001$; larger model probability error). Therefore, the current evidence does not support MS-CineMIL having a stable discriminative or probability advantage over the fixed 8-frame baseline. The corrected-input after field auditing and the majority-cines-positive sensitivity analysis further show that conclusions are sensitive to input fields and patient-level endpoint aggregation rules. The third-reviewer QC constitutes only a technical image quality audit. The main contribution of this paper is to establish and audit a patient-level dynamic ultrasound internal evaluation pipeline and to delineate the evidence boundaries of hierarchical aggregation under the current data and run conditions, rather than to demonstrate clinical utility or stable methodological superiority.

Declarations

Ethics Approval and Consent to Participate

This study is a secondary analysis of public de-identified data, reviewed and approved by the Biomedical Research Ethics Committee of Xuzhou Mining Group General Hospital (approval number: [2026] 080701; August 7, 2026). The original ethics approval and informed consent information for the public data acquisition are documented by the dataset publisher. This study involves no new subject recruitment, clinical intervention, or re-identification. This study was conducted in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable. This study is a secondary analysis of a public de-identified dataset, and the manuscript does not contain any individual person's data in any identifiable form.

Competing interests

The authors declare that they have no conflicts of interest.

Funding

This study received no external funding or research grants.

Author Contribution

Lei Jiao: Conceptualization; Methodology design; Software development and model implementation; Data preprocessing and manifest construction; Formal analysis (statistical analysis, paired bootstrap testing, calibration assessment); Original draft writing; Visualization. Zhen Zhou: Clinical data interpretation; Writing – review & editing. Hongsheng Yin: Methodology discussion; Software co-development; Model validation. Jinhu Ma: Data curation and collection; Investigation. Chuanhe Qian: Data curation and collection; Investigation. Tao Li: Resources; Investigation. Hao Zhang: Supervision; Project administration; Writing – review & editing. All authors have read and approved the final manuscript.

Acknowledgements

Not applicable.

Data Availability

The ThrombUS+ Ultrasound Dataset #1 Training Set (v1.0) is available via Zenodo (DOI: 10.5281/zenodo.17659415) under a CC BY 4.0 license. The audit manifests derived in this study (with manifest hashes), patient five-fold splits, all patient-level out-of-fold predictions, per-fold run signatures, and complete analysis code have been permanently archived via Zenodo with a DOI under a CC BY 4.0 license (archive DOI: 10.5281/zenodo.22043843; v2.7 release). The code for model training, patient-level cross-validation, out-of-fold aggregation, paired bootstrap, calibration, and sensitivity analyses has been version-controlled and audited in the research workspace, permanently archived via Zenodo with a DOI (10.5281/zenodo.22043843). The released materials include all scripts for the prespecified compared models, ablations, field-completed re-runs, label aggregation, and same-environment controlled comparisons, as well as the run signature records in each fold report (software versions, GPU model, manifest hash, and run identifier), enabling readers to independently verify the between-run variance audit results.

Code Availability

The code for model training, patient-level cross-validation, out-of-fold aggregation, paired bootstrap, calibration, and sensitivity analyses has been version-controlled and audited in the research workspace, permanently archived via Zenodo with a DOI (10.5281/zenodo.22043843). The released materials include all scripts for the prespecified compared models, ablations, field-completed re-runs, label aggregation, and same-environment controlled comparisons, as well as the run signature records in each fold report (software versions, GPU model, manifest hash, and run identifier), enabling readers to independently verify the between-run variance audit results.

References

1. Wendelboe AM, Raskob GE. Global burden of thrombosis: epidemiologic aspects. *Circ Res*. 2016;118(9):1340–7. 10.1161/CIRCRESAHA.115.306841.
2. Di Nisio M, van Es N, Büller HR. Deep vein thrombosis and pulmonary embolism. *Lancet*. 2016;388(10063):3060–73. 10.1016/S0140-6736(16)30514-1.
3. Bhatt M, Braun C, Patel P, et al. Diagnosis of deep vein thrombosis of the lower extremity: a systematic review and meta-analysis of test accuracy. *Blood Adv*. 2020;4(7):1250–64. 10.1182/bloodadvances.2019000960.
4. Pomero F, Dentali F, Borretta V, et al. Accuracy of emergency physician-performed ultrasonography in the diagnosis of deep-vein thrombosis: a systematic review and meta-analysis. *Thromb Haemost*. 2013;109(1):137–45. 10.1160/TH12-07-0473.
5. Needleman L, Cronan JJ, Lilly MP et al. Ultrasound for lower extremity deep venous thrombosis: multidisciplinary recommendations from the Society of Radiologists in Ultrasound consensus conference. *Circulation*. 2018;137(14):1505–1515. 10.1161/CIRCULATIONAHA.117.030687
6. Shen YT, Chen L, Yue WW, Xu HX. Artificial intelligence in ultrasound. *Eur J Radiol*. 2021;139:109717. 10.1016/j.ejrad.2021.109717.
7. Kainz B, Heinrich MP, Makropoulos A, et al. Non-invasive diagnosis of deep vein thrombosis from ultrasound imaging with machine learning. *npj Digit Med*. 2021;4:137. 10.1038/s41746-021-00503-7.
8. Oppenheimer J, Mandegaran R, Staabs F, et al. Remote expert DVT triaging of novice-user compression sonography with AI-guidance. *Ann Vasc Surg*. 2024;99:272–9. 10.1016/j.avsg.2023.08.022.
9. Speranza G, Mischkewitz S, Al-Noor F, Kainz B. Value of clinical review for AI-guided deep vein thrombosis diagnosis with ultrasound imaging by non-expert operators. *npj Digit Med*. 2025;8:135. 10.1038/s41746-025-01518-0.
10. van Sloun RJG, Cohen R, Eldar YC. Deep learning in ultrasound imaging. *Proc IEEE*. 2020;108(1):11–29. 10.1109/JPROC.2019.2932116
11. Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR. 2018;80:2127–2136.
12. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301–9. 10.1038/s41591-019-0508-1.
13. Schlemper J, Oktay O, Schaap M, et al. Attention gated networks: learning to leverage salient regions in medical images. *Med Image Anal*. 2019;53:197–207. 10.1016/j.media.2019.01.012.
14. Shao Z, Bian H, Chen Y et al. TransMIL: transformer based correlated multiple instance learning for whole slide image classification. In: *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. 2021.
15. Tu Y, Lei H, Long W, Yang Y. HAMIL: hierarchical aggregation-based multi-instance learning for microscopy image classification. *Pattern Recognit*. 2023;136:109245.

- 10.1016/j.patcog.2022.109245.
16. Xiong C, Chen H, Sung JJY, King I. Diagnose like a pathologist: transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI). 2023.
 17. Didaskalou S, Portokallidis N, Tzatzimaki K, Liapi M, Moustakidis N, Moustakidis T et al. A dataset for benchmarking machine learning models for autonomous deep vein thrombosis detection based on compression ultrasound videos. In: Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2026), vol 4: HEALTHINF. SciTePress; 2026:853–862. ISBN: 978-989-758-802-0.
 18. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell.* 2021;3(3):199–217. 10.1038/s42256-021-00307-0.
 19. Howard A, Sandler M, Chu G et al. Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:1314–1324. 10.1109/ICCV.2019.00140
 20. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev.* 1950;78(1):1–3. 10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2.
 21. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap.* New York: Chapman & Hall/CRC; 1993.
 22. Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola AJ, Bartlett P, Schölkopf B, Schuurmans D, editors. *Advances in Large Margin Classifiers.* MIT Press; 1999. pp. 61–74.
 23. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol.* 2016;74:167–76. 10.1016/j.jclinepi.2015.12.005.
 24. Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271.* 2018.
 25. Bouthillier X, Delaunay P, Bronzi M, et al. Accounting for variance in machine learning benchmarks. *Proc Mach Learn Syst.* 2021;3:747–69.
 26. Picard RR, Cook RD. Cross-validation of regression models. *J Am Stat Assoc.* 1984;79(387):575–83. 10.1080/01621459.1984.10478083.
 27. Nadeau C, Bengio Y. Inference for the generalization error. *Mach Learn.* 2003;52(3):239–81. 10.1023/A:1024068626366.
 28. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* 1998;10(7):1895–923. 10.1162/089976698300017197.
 29. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated ROC curves: a nonparametric approach. *Biometrics.* 1988;44(3):837–45. 10.2307/2531595.
 30. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17(1):230. 10.1186/s12916-019-1466-7.

31. Wells PS, Anderson DR, Bormanis J, et al. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. Lancet. 1997;350(9094):1795–8. 10.1016/S0140-6736(97)08140-3.
32. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.
33. Mongan J, Moy L, Kahn CE Jr.. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell. 2020;2(2):e200029. 10.1148/ryai.2020200029.
34. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8. 10.7326/M18-1376.
35. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691–8. 10.1136/heartjnl-2011-301247.

Figures

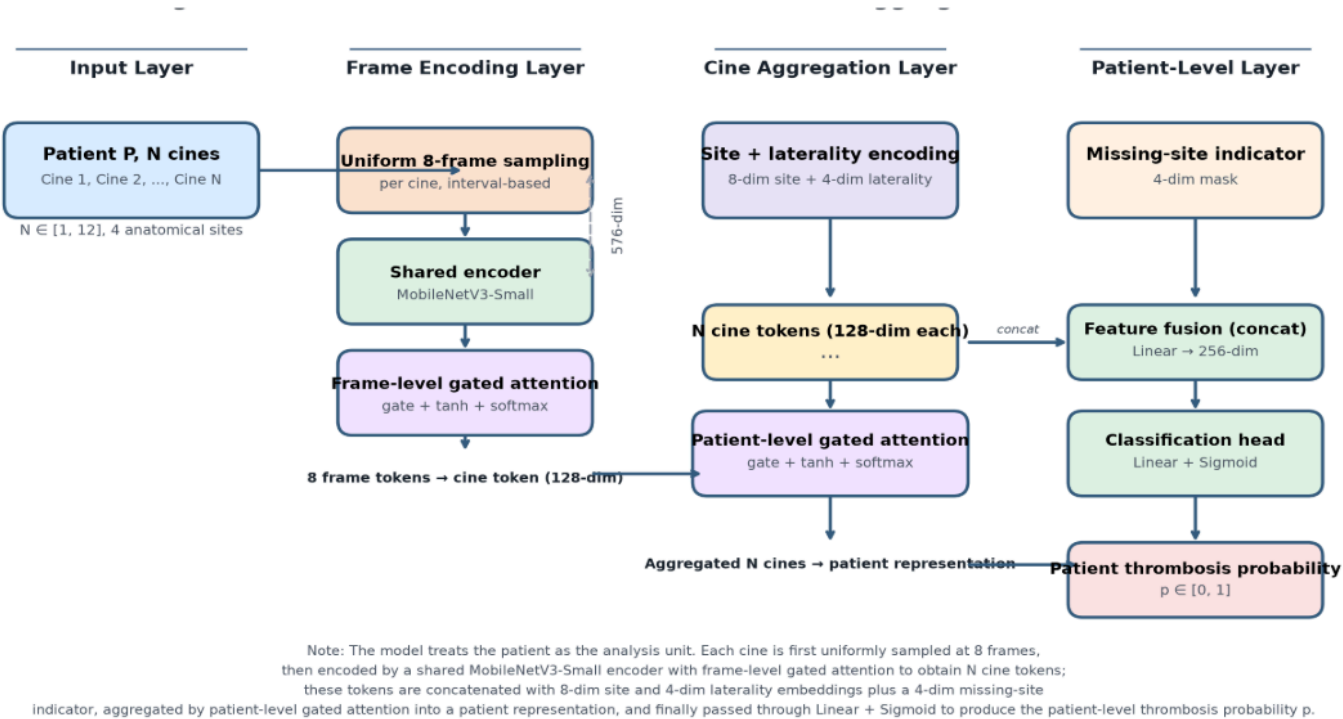


Figure 1

Schematic overview of the MS-CineMIL hierarchical aggregation architecture

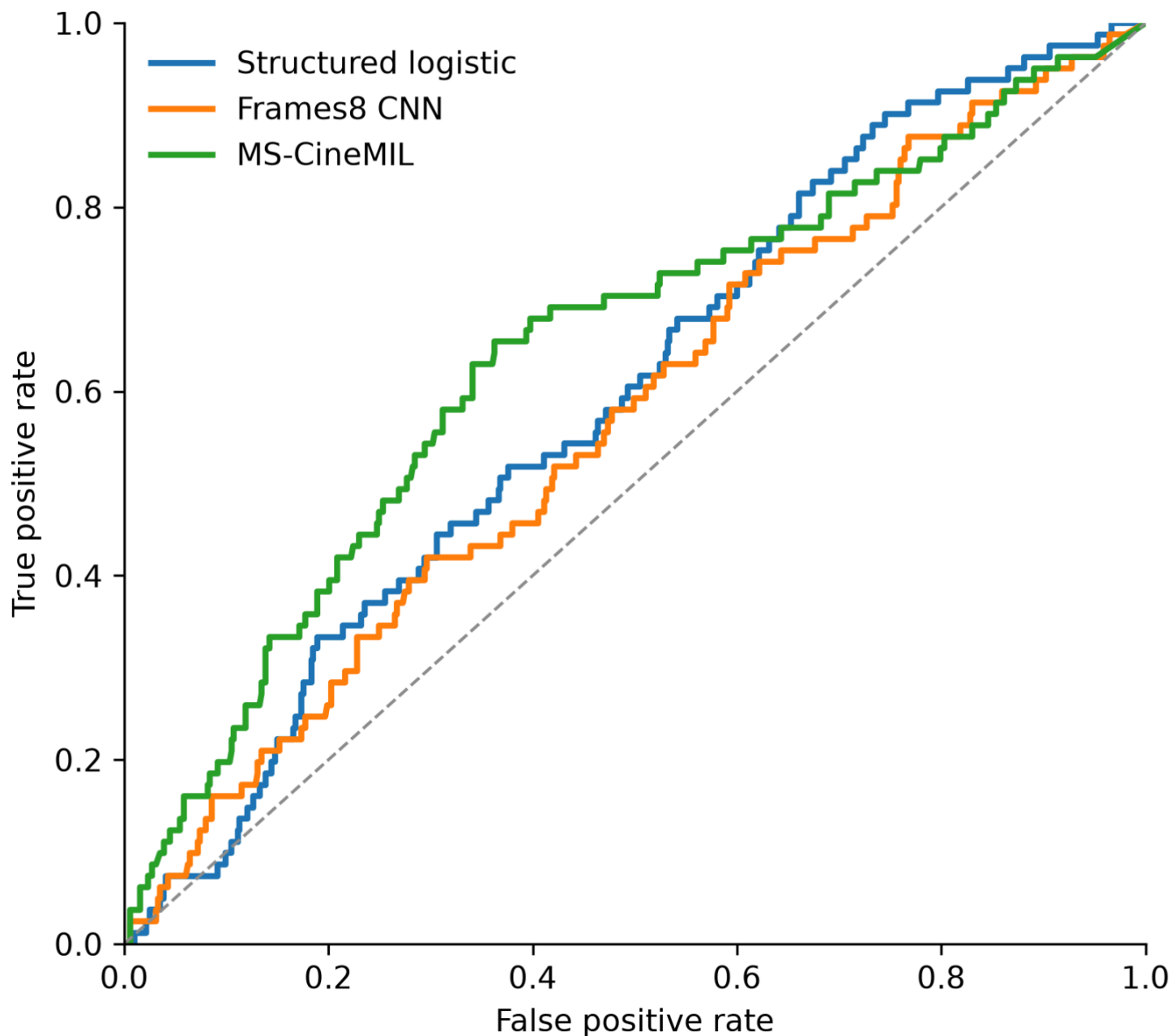


Figure 2

Patient-level out-of-fold ROC curves of the three prespecified compared models. Source: patient-level five-fold out-of-fold (OOF) predictions of the three prespecified compared models from the V1 frozen analysis; this figure does not include corrected-input re-runs or same-environment controlled comparisons, and no external validation data are used.

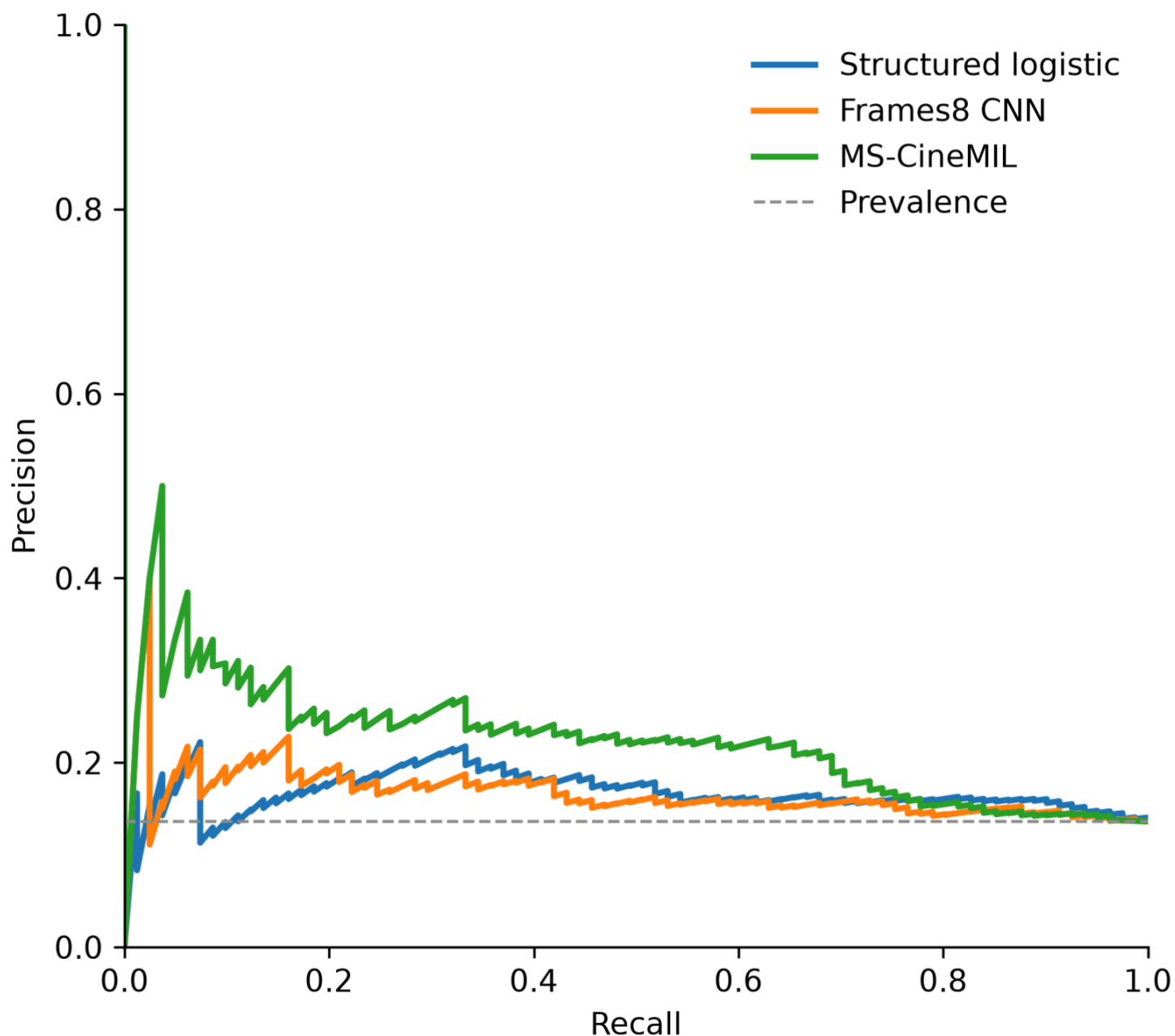


Figure 3

Patient-level out-of-fold PR curves of the three prespecified compared models. Source: patient-level five-fold out-of-fold (OOF) predictions of the three prespecified compared models from the V1 frozen analysis; this figure does not include corrected-input re-runs or same-environment controlled comparisons, and no external validation data are used.

MS-CineMIL Core Ablations: Paired 95% Bootstrap CIs

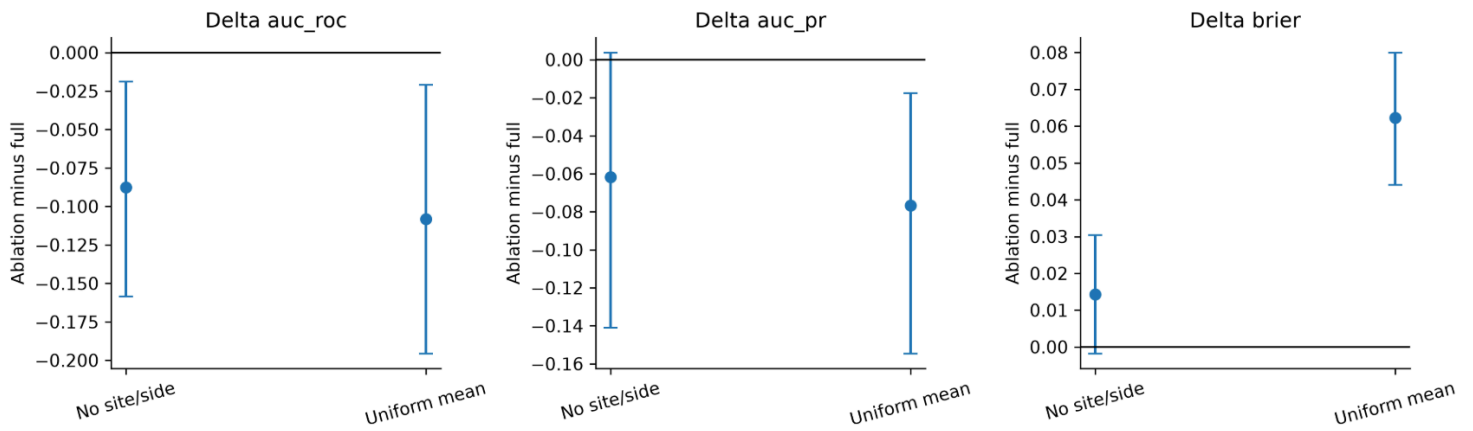


Figure 4

Performance changes of core ablations relative to the complete MS-CineMIL with 95% confidence intervals. Source: patient-level five-fold out-of-fold (OOF) predictions of the ablation models and the complete MS-CineMIL from the V1 frozen analysis; this figure does not include corrected-input re-run results, and no external validation data are used.

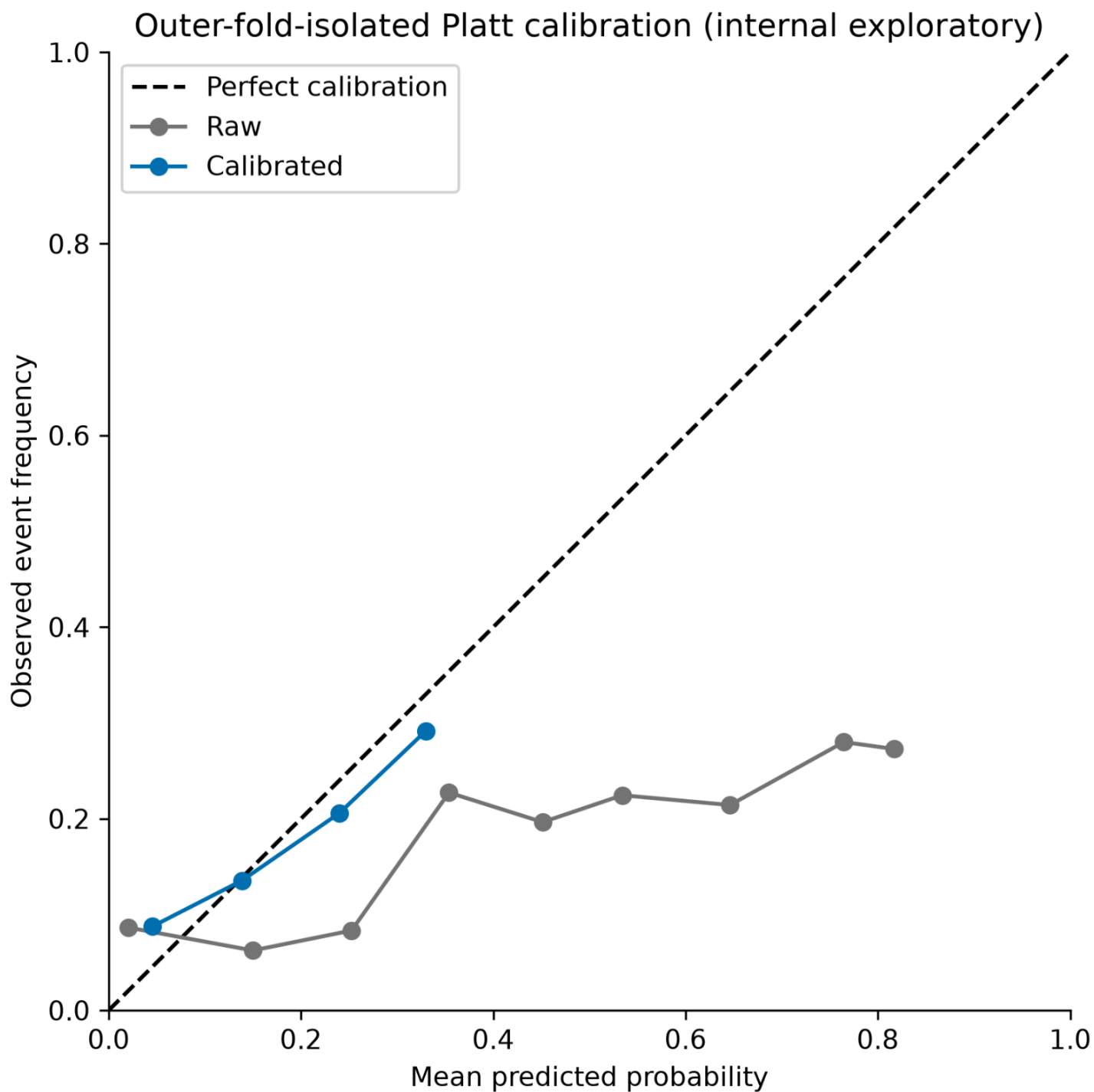


Figure 5

Reliability curves of MS-CineMIL before and after exploratory Platt calibration. Source: MS-CineMIL patient-level five-fold out-of-fold (OOF) predictions from the V1 frozen analysis; no external validation data are used.

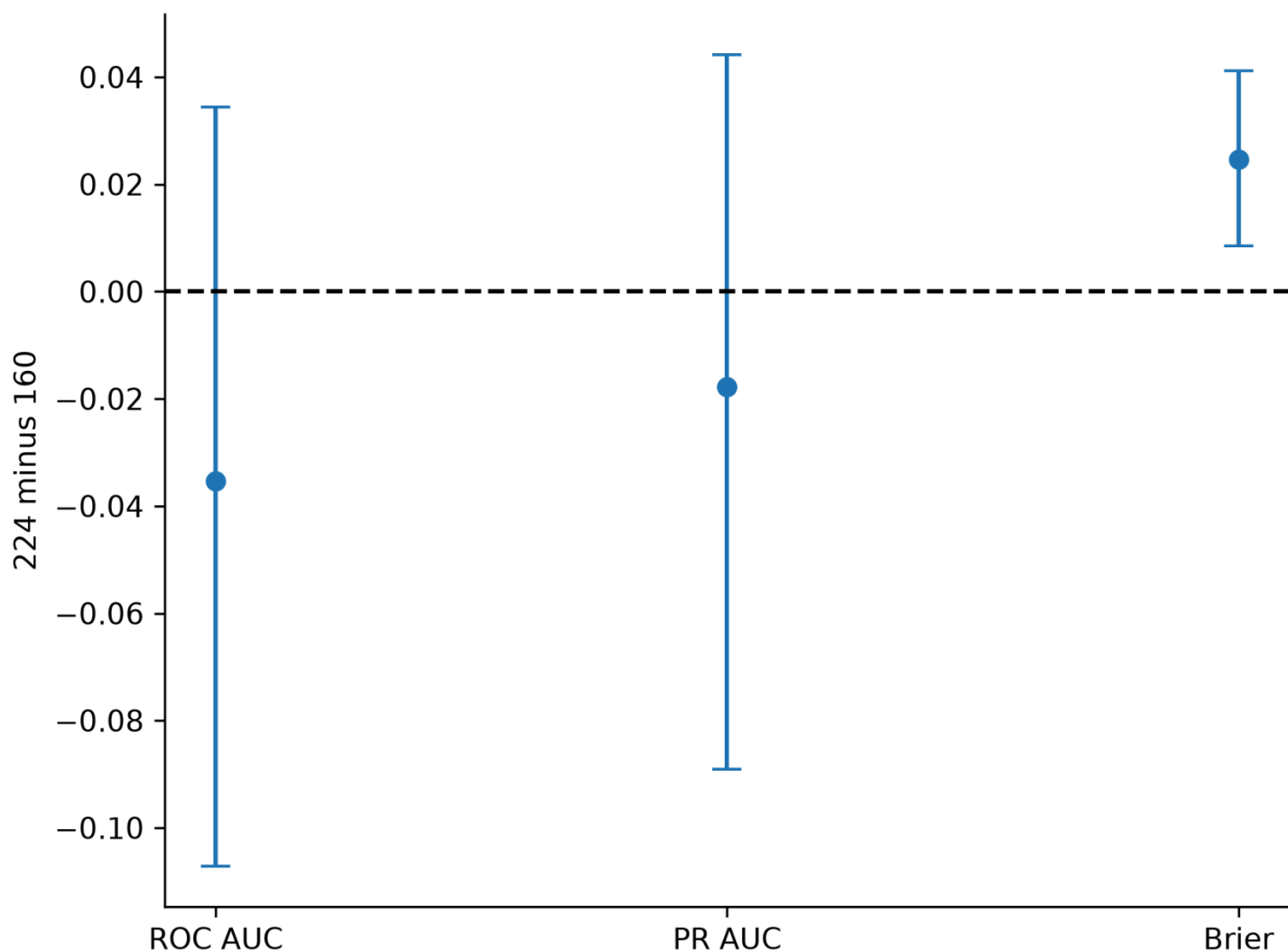


Figure 6

Patient-level paired performance differences of 224×224 relative to the frozen 160×160 main model. Source: patient-level five-fold out-of-fold (OOF) predictions from the prespecified resolution sensitivity analysis (executed under the V1 frozen analysis framework and the same patient folds); no external validation data are used.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterial.docx](#)