

Does Probability Calibration Survive a Change of OCT Scanner? A Leakage-Free Study of Calibration Transfer Across Manufacturers

Tanha Asfika Jaman
Nanjing University of Information Science and
Technology (NUIST)
Nanjing, China

Akter Mst Lovely
Nanjing University of Information Science and
Technology (NUIST)
Nanjing, China

Iftee Shekh Iftesham
Nanjing University of Information Science and
Technology (NUIST)
Nanjing, China

Mostafa Farzana
Nanjing University of Information Science and
Technology (NUIST)
Nanjing, China

ABSTRACT

Retinal optical coherence tomography (OCT) classifiers are usually reported with calibration measured on the same scanner on which they were trained. This study measures what happens to their predicted probabilities on a different manufacturer's scanner. Three-class classifiers (AMD, DME, normal) are trained on 65,585 patient-disjoint Heidelberg Spectralis B-scans (Kermany) and evaluated on an unseen Optovue Avanti RTVue XR cohort (OCTDL) across eleven runs, two backbones and two resizing policies. The data audit finds two leakage channels: Kermany patient identifiers are global, being shared between class folders 501 to 2,597 times more often than chance, and in OCTDL 47 pairs of pixel-identical images carry different patient identifiers. Splits are therefore built on connected components of (same patient) OR (near-duplicate). The resizing policy matters most for accuracy: square resizing raises target balanced accuracy from 0.8821 to 0.9304 and lowers uncalibrated expected calibration error (ECE) from 0.0316 to 0.0181 relative to letterboxing, a reduction comparable to that of the best source-only calibrator. Source-only temperature scaling lowers target ECE in 10 of 11 runs (95% bootstrap interval excluding zero in 6); calibrating on source data degraded by a physically parameterised OCT acquisition model beats a generic corruption family of matched severity in only 5 of 11 runs; and input-conditioned temperature scaling loses to a single global temperature in 10 of 11 runs. A referral threshold promising 0.5% error on source data yields 1.19% to 1.59% error on the new scanner.

General Terms

Algorithms, Experimentation, Measurement, Reliability

Keywords

Optical coherence tomography, probability calibration, domain shift, data leakage, temperature scaling, retinal imaging, reproducibility

1. INTRODUCTION

A retinal OCT screening model does not output a diagnosis. It outputs a probability, and a clinic converts that probability into a triage decision by thresholding it. The usefulness of such a system therefore depends on the probability being trustworthy, not only on the arg-max being correct. An accurate but over-confident model refers too few uncertain cases for review, which accuracy alone cannot reveal.

Recent OCT studies report calibration routinely, adding temperature scaling and reliability diagrams [1, 2], but report it on data from the training cohort. Cross-scanner OCT studies concentrate on discrimination [3, 4]. The most relevant cross-scanner study reports that expected calibration error “remains elevated under severe dataset shifts” and that its conformal sets need site-specific calibration with labelled target data [3]; it does not compare post-hoc calibrators that could be fitted before deployment. Large OCT foundation models [5] improve representations but do not answer whether probabilities can be trusted on a new device.

On the first day of deployment at a site with a different scanner there is no labelled target data. The first question is therefore what happens to the probabilities, and what can be done in advance using source data alone.

Target-free calibration methods expose the model to simulated degradation of source data [6, 7] built from generic corruptions such as noise, blur and JPEG [8]. None is a physical OCT effect: OCT degradation arises from multiplicative speckle, axial point-spread broadening, depth-dependent sensitivity roll-off, lateral sampling and vendor log compression. How well synthetic corruptions predict real shift depends on how closely the simulated mechanism matches the real one (per-corruption correlation 0.270 to 0.867), but this was measured for segmentation robustness on driving scenes, not for calibration [9]. The second question is therefore whether the physical faithfulness of the simulated shift determines whether a source-only calibrator transfers to a real, unseen device.

Contributions. Each contribution is stated at the strength the evidence supports, and its kind is labelled.

- *Dataset audit finding.* Two leakage channels in widely used OCT corpora are identified and quantified; each survives the per-subject splitting prescribed by the standard reference on OCT leakage [10] (Section 3).
- *Empirical finding.* Source-only post-hoc calibrators are compared across a real OCT manufacturer change under a patient- and duplicate-disjoint protocol with eleven runs. The targeted search in Section 2 did not identify a prior study making this comparison (Section 6).
- *Engineering artefact.* A physically parameterised OCT acquisition-shift family with seven axes, applied in the

linearised intensity domain, is implemented as a reusable software module (Section 4.2).

- *Negative finding.* Physical faithfulness of the simulated shift does not reliably improve calibration transfer, while the image-resizing policy changes accuracy, which no calibrator can, and reduces ECE about as much as the best source-only calibrator (Sections 6.2 and 6.5).

No new network architecture is proposed and no state-of-the-art accuracy claim is made. The classifiers are standard ImageNet-pretrained backbones fine-tuned with a standard recipe and are used as measurement instruments rather than as contributions. Figure 1 summarises the complete study pipeline.

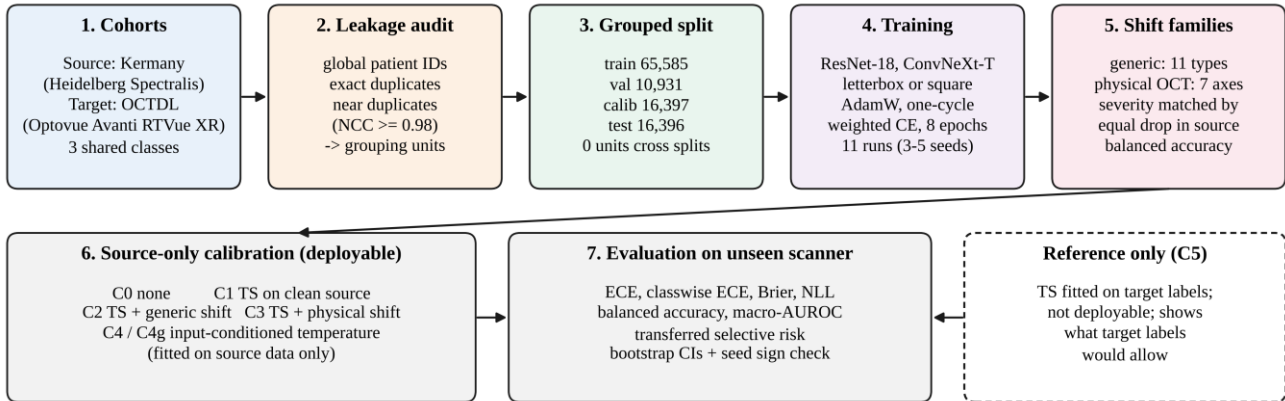


Figure 1. Overview of the study. Steps 1–5 build a leakage-free source model and two matched families of simulated shift; step 6 fits six calibration conditions on source data only; step 7 measures them on the unseen scanner. C5 (dashed) uses target labels and serves only as a non-deployable reference.

2. RELATED WORK

Post-hoc calibration. Temperature scaling [11] is the default post-hoc calibrator and preserves accuracy, since dividing all logits by one positive scalar cannot change the arg-max. Calibration degrades under dataset shift in general [12]. Transferable calibration [13] needs an unlabelled target sample; local temperature scaling [14] varies the temperature spatially for segmentation; and sample-dependent adaptive temperature scaling [15] conditions the temperature on the input. The input-conditioned calibrator evaluated here is an adaptation of the latter to the cross-device setting, not a new method.

Target-free calibration under shift. Corruption-based calibration [6] fits the calibrator on synthetically degraded source data, and frequency-aware gradient rectification [7] achieves target-agnostic calibration at training time. Both are validated on generic corruption families and generic benchmarks.

Cross-vendor calibration in medical imaging. CalSAM [16] calibrates brain MRI segmentation across a Siemens-to-GE vendor shift without target data, using a Fisher-information penalty and a confidence-misalignment penalty. It is the closest prior work in spirit but differs in modality, task and mechanism: it is a training-time regulariser, whereas the present study asks whether the fidelity of a simulated shift governs post-hoc calibration transfer.

OCT reliability and cross-scanner OCT. Two 2026 studies add post-hoc calibration to a modern backbone and report low ECE on test data from the training cohort [1, 2]. NA-DyCNN [3] makes the representation speckle-robust, evaluates zero-shot transfer across three cohorts and reports that ECE remains elevated under severe shift, with a conformal fallback that needs labelled target data. Lopukhova et al. [4] include cross-scanner verification and report an ECE of $2.1 \pm 0.4\%$ on a held-out fold of the training data. Calibration degradation under OCT scanner shift is therefore known; what these studies do not report is how source-only post-hoc calibrators compare

after a manufacturer change, or whether the kind of simulated shift used to fit them matters.

Leakage and corruption benchmarks. Tampu et al. [10] showed that per-image instead of per-subject splitting inflates the Matthews correlation coefficient by 0.08 to 0.43 across OCT datasets, including 0.12 and 0.07 for versions 2 and 3 of Kermany. The present audit finds two further channels that per-subject splitting does not close. ImageNet-C [8] defines the generic corruption family, and MedMNIST-C [17] applies generic corruptions to MedMNIST, where OCT appears only as 28×28 thumbnails. The literature search (OCT calibration, calibration under shift, cross-vendor calibration and corruption benchmarks) did not identify a study that compares source-only calibrators, including calibrators fitted on simulated acquisition shift, across a real OCT manufacturer change; novelty is claimed only at that strength.

3. DATA AND DATA INTEGRITY

3.1 Cohorts and Label Space

Source. Kermany [18], Mendeley Data v3 (DOI 10.17632/rschjbr9sj.3), Heidelberg Spectralis, CC BY 4.0, 109,309 OCT images; the archive SHA-256 was verified against the published hash (this record is larger than the common 84,495-image “OCT2017” subset). **Target.** OCTDL [19], Mendeley Data v4 (DOI 10.17632/sncdhf53xc.4), Optovue Avanti RTVue XR, CC BY 4.0, SHA-256 verified, 2,064 images and 820 distinct patient identifiers computed from the released file (the publication reports 821).

Three classes form the clinically coherent intersection: AMD (Kermany CNV + DRUSEN; OCTDL AMD), DME and NORMAL. Kermany CNV is neovascular AMD and DRUSEN is early or intermediate AMD, so their union spans the AMD spectrum, which OCTDL’s AMD class also spans. OCTDL’s ERM, RVO, VID and RAO classes (354 images) have no source counterpart and are excluded. The three-class target subset has 1,710 images from 637 patients (AMD 1,231, DME 147, NORMAL 332). De-duplication keeps one image (the largest file) per near-duplicate component ($NCC \geq 0.98$,

Section 3.3), leaving 1,606 images (AMD 1,154, DME 139, NORMAL 313) from 618 patients in 575 grouping units; the 19 patients who drop out have all their images represented by a near-duplicate kept under another patient identifier. The pairing is a simultaneous manufacturer, site and population shift; this is the realistic deployment condition and also a limitation, because effects cannot be attributed to the manufacturer alone.

3.2 Kermany Patient Identifiers Are Global

The folder layout invites keying a patient-disjoint split on (class, patient_id), which is incorrect. Identifiers span 1,384 to 9,997,680, a range holding 5,420 identifiers at a density of 0.054%, so two class folders would share an identifier by chance well under once. The observed number of shared identifiers exceeds this chance expectation by $501\times$ (CNV–NORMAL) to $2,597\times$ (CNV–DRUSEN): 222 identifiers are shared between CNV and DRUSEN, 107 between CNV and DME, 86 between DME and DRUSEN, 253 between DME and NORMAL, 203 between DRUSEN and NORMAL and 175 between CNV and NORMAL. Independently, the archive’s own train/test split is exactly disjoint under the raw identifier (4,785 train, 635 test, 0 overlap). The identifiers therefore identify people, and 787 patients legitimately carry images under several diagnoses. Keying on (class, identifier) produces 6,332 apparent patients instead of 5,420 and splits 637 patients and 24,398 images, 22.3% of the corpus, across folds.

3.3 Duplicate Content and the Grouping Unit

Kermany. 15,008 images (13.7%) are exact pixel duplicates of another image, in 7,117 groups; 269 groups span two original class folders, but only 82 span two different *study* labels, because CNV and DRUSEN both map to AMD. Those 82 groups contain 177 images (0.16%) and are genuine label noise. **OCTDL.** 55 image pairs are element-wise identical after decoding (verified by array comparison, not by hash alone), and 47 of them are filed under two different patient identifiers, so a perfectly patient-disjoint split can still place the same image on both sides. Near duplicates are detected by normalised cross-correlation (NCC) of intensity-standardised thumbnails (64×64 for OCTDL, 32×32 for Kermany),

$$\text{NCC}(a, b) = \sum_i (a_i - \bar{a})(b_i - \bar{b}) / [\|a - \bar{a}\| \|b - \bar{b}\|], \quad (1)$$

whose OCTDL null distribution reaches 0.795 at the 99.9th percentile against at least 0.99999 for confirmed duplicates; 13.2% of the 2,064 OCTDL images lie in a redundancy group at $\text{NCC} \geq 0.98$.

Grouping unit. An undirected graph is built whose nodes are images and whose edges join two images if they share a patient identifier or have $\text{NCC} \geq 0.98$. Each connected component is one grouping unit, used both for splitting and for the cluster bootstrap. Exact duplicates have $\text{NCC} = 1$ and are absorbed into the same component, so the 82 contradictory-label groups are confined to one split. The 177 contradictory images are retained as label noise, because removing them would alter class balance and their effect is shared identically by every condition. Table 1 verifies every leakage channel against the realised split assignment rather than against the split code.

Table 1. Leakage channels, verified on the realised split

Channel	Count	Status
Patients in more than one split	0	closed
Exact-duplicate groups spanning splits	0	closed
Grouping units spanning splits	0	closed
Images pixel-identical across cohorts	0	closed
Contradictory-label groups spanning splits	0	confined

The source data are split by grouping unit into train 65,585, validation 10,931, calibration 16,397 and test 16,396 images over 4,920 units, with class proportions matching within 0.01 percentage points across all four. The calibration split is used only to fit calibrators and to fix operating thresholds, never for model selection.

4. METHODOLOGY

4.1 Classifiers and Training Protocol

Two ImageNet-pretrained backbones are fine-tuned: ResNet-18 (5 seeds) and ConvNeXt-Tiny (3 seeds) under aspect-preserving letterboxing, and ResNet-18 under naive square resizing (3 seeds), giving eleven runs and three configurations. Each B-scan is converted to greyscale, resized to 224×224 by the configuration’s policy and standardised per image to zero mean and unit variance. Letterboxing scales the longer side to 224 pixels and pads the shorter side with zeros; square resizing stretches both axes independently. Augmentation is limited to a lateral flip (probability 0.5) and a small translation (probability 0.5; up to 5% laterally and 3% axially). Vertical flips are excluded because the depth axis has a fixed anatomical orientation.

Training minimises class-weighted cross-entropy,

$$\mathcal{L} = -(1/N) \sum_n w_{c(n)} \log p_{c(n)}(x_n), \quad w_c = N / (K N_c), \quad (2)$$

where $c(n)$ is the label of image x_n , $K = 3$ and N_c is the training count of class c , giving weights 0.79 (AMD), 3.14 (DME) and 0.71 (NORMAL). Optimisation uses AdamW (peak learning rate 3×10^{-4} , weight decay 10^{-4}) with a one-cycle schedule for 8 epochs and mixed precision on a single NVIDIA RTX 4060 Ti (8 GB). The checkpoint with the lowest validation loss is selected; no target image is seen during training or selection. Figure 2 shows that training loss falls smoothly in all eleven runs and that validation loss reaches its minimum between epochs 2 and 7, after which the selected checkpoint is frozen.

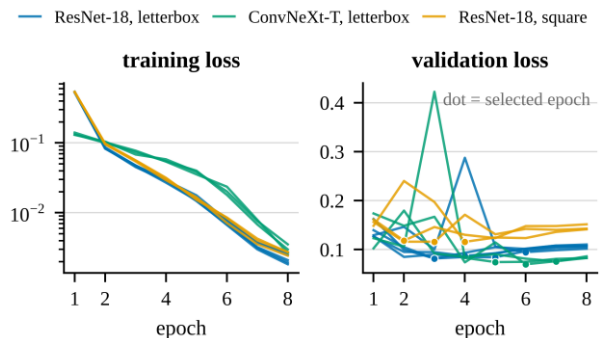


Figure 2. Training and validation loss for all eleven runs. Dots mark the epoch selected by minimum validation loss.

4.2 Two Families of Simulated Shift

Generic family (control). The primitives used in the target-free calibration literature [8]: Gaussian, shot and impulse noise; defocus, motion and zoom blur; contrast, brightness, elastic,

pixelate and JPEG, each at five severity levels applied to the displayed 8-bit image.

Physical family (treatment). Seven axes, each a mechanism by which two spectral-domain OCT scanners differ. The central design decision is that all operations act on linearised backscatter intensity rather than on displayed pixel values, because a rendered B-scan is log-compressed while speckle is multiplicative, detector noise additive and the axial point-spread function convolutional on linear intensity. A displayed grey level $g \in [0, 255]$ is mapped to linear intensity by inverting an assumed log compression with dynamic range $D = 45$ dB,

$$I = F \cdot 10^{(D/10)(g/255)}, \quad F = 10^{-D/10}, \quad (3)$$

and mapped back after degradation. The axes are: (1) multiplicative speckle $I' = I \cdot s$ with $s \sim \text{Gamma}(L, 1/L)$ for L effective looks, spatially correlated over the speckle size; (2) axial point-spread broadening, a Gaussian blur along depth only; (3) lateral sampling, decimation and re-interpolation along the lateral axis only; (4) sensitivity roll-off, a depth-dependent gain $I'(z) = I(z) \cdot 10^{-\alpha z/10}$ for normalised depth z ; (5) a raised noise floor, additive Gaussian noise of standard deviation $F \cdot 10^{r/10}$ on linear intensity; (6) axial scale and B-scan tilt; and (7) a different display dynamic range and tone curve. The vendor log mapping is not published and cannot be recovered from a rendered image, so a negative result for this family is evidence about this model of OCT image formation, not about OCT physics in general.

4.3 Functional Severity Matching

Comparing the families is meaningful only at comparable strength, and no neutral image statistic exists on which to match them, because any such statistic is a property that one family models and the other does not. Severity is therefore matched functionally. For each run and family, the balanced accuracy $BA(s)$ of the source model is measured on the degraded source validation split at integer severities $s = 0, \dots, 5$, where each image receives one randomly chosen axis of the family, and the matched severity is

$$s^* = \min \{ s : BA(s) \leq BA(0) - \Delta \}, \quad (4)$$

obtained by linear interpolation between bracketing integer levels (a fractional severity is realised by sampling the two neighbouring levels with the corresponding probabilities). The target drop Δ is the largest that both families can reach: 5 points for ResNet-18 but only 2 points for ConvNeXt-Tiny, because the generic family cannot produce a 5-point drop on ConvNeXt-Tiny at any severity (Figure 3). This asymmetry is itself informative; a fixed target would have silently removed the generic condition. For example, ResNet-18 seed 0 is matched at $s^* = 3.77$ (generic) and 3.06 (physical).

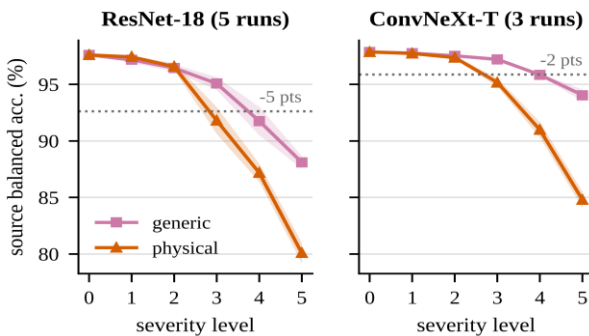


Figure 3. Functional severity matching: source balanced accuracy versus severity for both shift families (mean $\pm$ SD over runs). The dotted line marks the target drop at which the two families are compared.

4.4 Calibration Conditions

Let $z(x) \in \mathbb{R}^K$ be the logits of image x . Scalar temperature scaling outputs $p(x) = \text{softmax}(z(x)/T)$, with T fitted by minimising negative log-likelihood (NLL) on a fitting set S ,

$$T^* = \arg\min_{T>0} -\sum_{(x,y) \in S} \log \text{softmax}_y(z(x)/T), \quad (5)$$

solved with L-BFGS over $\log T$. Seven conditions are compared: **C0**, raw softmax; **C1**, T fitted on the clean source calibration split; **C2**, T fitted on the union of the clean calibration split and a copy degraded by the generic family at s^* ; **C3**, the same with the physical family; **C4**, an input-conditioned temperature $T(x) = \exp(\text{clip}(g(f(x)), -3, 3))$, where $f(x)$ is the 512- or 768-dimensional penultimate feature and g is a 64-unit one-hidden-layer perceptron trained with Adam (learning rate 10^{-3} , batch 512) on the physical union set; **C4g**, the same on the generic union set; and **C5**, T fitted on the labelled target set. Only C5 uses target labels; it is a non-deployable reference and, because it minimises NLL rather than ECE, it is not a strict lower bound on ECE. All conditions preserve each image's arg-max, so accuracy is identical across C0–C5 within a run.

4.5 Evaluation Metrics

The headline calibration metric is adaptive (equal-mass) top-label ECE with $M = 15$ bins B_m of equal size,

$$\text{ECE} = \sum_{m=1}^M (|B_m|/N) \cdot |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (6)$$

where acc and conf are the mean correctness and mean top-label confidence in the bin. Equal-width ECE, classwise ECE (the one-vs-rest ECE averaged over classes), top-label Brier score with Murphy's calibration-refinement decomposition, NLL, the area under the risk-coverage curve (AURC) and its excess (E-AURC), balanced accuracy, macro-AUROC, macro-F1 and Cohen's kappa are also computed. Degradation under the scanner change is expressed as the error ratio

$$R(m) = e_{\text{tgt}}(m) / e_{\text{src}}(m), \quad (7)$$

where $e(m)$ is the metric itself for error-like metrics (ECE, Brier, NLL, E-AURC) and $1 - m$ for discrimination metrics (AUROC, balanced accuracy), so that all ratios are comparable. The **transferred operating point** measures clinical consequence: a confidence threshold τ is fixed on the clean source calibration split so that the selective error among retained cases equals a promised risk $r \in \{0.25\%, 0.5\%, 1\%\}$, and the same τ is applied unchanged on the target.

4.6 Statistical Protocol

Two kinds of uncertainty are reported and kept separate. Seed variation is summarised as mean $\pm$ SD across runs, where SD is the population standard deviation (divisor equal to the number of runs). Sampling uncertainty within one trained model is estimated by a cluster bootstrap over grouping units (1,000 resamples, percentile 95% intervals); paired comparisons share resamples, and their two-sided p-values are bootstrap estimates. Because a per-run interval says nothing about whether an effect survives retraining, a claim is accepted only if it is also consistent in sign across runs, and an effect smaller than its across-seed SD is not reported as a finding. Configurations are analysed separately and never pooled. Because scalar temperature scaling is rank-preserving, C0–C3 share an identical risk-coverage curve, and AURC is used only to characterise the domain gap. All numbers are produced by the analysis scripts from saved logits, and a verification script checks the reported values against the result files.

5. VERIFYING THAT THE DEVICE SHIFT IS REAL

Before calibration is studied, the cohorts are tested for separability by acquisition statistics with geometry controlled. Forty-one intensity, speckle, spectral and roll-off statistics computed on native-resolution 128×128 patches (1,074 source and 1,066 target images) separate the devices at AUROC 0.9996 (gradient boosting, patient-grouped 5-fold cross-validation). The twenty structural statistics that no global tone curve can produce still give 0.9947, four anisotropy features alone give 0.914, and a group-level label permutation returns 0.5046 (maximum 0.5437). The largest univariate effects are axial-minus-lateral spectral slope (AUROC 0.853, $d = +1.44$), axial high-frequency energy (0.852, $d = +1.50$) and high-frequency anisotropy (0.849, $d = +1.45$): the devices differ mainly in axial versus lateral resolution. The device shift is therefore real and characterisable.

6. RESULTS

Results are reported for three configurations: ResNet-18 with letterboxing (R18-LB, 5 runs), ConvNeXt-Tiny with letterboxing (CNX-LB, 3 runs) and ResNet-18 with square resizing (R18-SQ, 3 runs). Run counts are stated as “k of 11” over all runs unless a subset is named.

6.1 In-Domain Performance and Generalisation Gap

In every run the checkpoint selected on validation loss is the validation optimum (11 of 11), and validation loss rebounds only modestly afterwards (mean 0.0205, 0.0104 and 0.0285 for R18-LB, CNX-LB and R18-SQ), the mild late-epoch overfitting that checkpoint selection absorbs. Validation accuracy and held-out source-test accuracy differ on average by -0.0042 , -0.0028 and $+0.0014$, so the in-domain generalisation gap is essentially zero. Source-test balanced accuracy is 0.9786, 0.9803 and 0.9532, and target balanced accuracy is 0.8821, 0.9157 and 0.9304, a cross-domain drop of 9.6, 6.5 and 2.3 points; the degradation studied here is therefore attributable to the domain change rather than to overfitting.

Figure 4 shows where the errors occur on the target scanner. AMD and NORMAL recall remain between 94.6% and 98.0% in every configuration, whereas DME recall falls to 71.5% (R18-LB), 80.6% (CNX-LB) and 88.5% (R18-SQ), against 90.7% to 97.1% on the source test split. The dominant failure is DME predicted as AMD (21.2%, 15.6% and 10.6% of target DME images), consistent with DME being the minority class (139 target images) and with both diseases presenting retinal thickening and hyper-reflective deposits.

	R18-LB (5 runs)			CNX-LB (3 runs)			R18-SQ (3 runs)		
	AMD	DME	NOR	AMD	DME	NOR	AMD	DME	NOR
AMD	95.4	0.1	4.4	96.2	0.5	3.3	96.1	1.9	2.0
DME	21.2	71.5	7.3	15.6	80.6	3.8	10.6	88.5	1.0
NOR	2.1	0.2	97.7	1.6	0.4	98.0	2.8	2.7	94.6
	predicted			predicted			predicted		

Figure 4. Row-normalised confusion matrices on the unseen scanner (percent, averaged over runs). The dominant error is DME classified as AMD; it shrinks under square resizing (NOR = NORMAL).

6.2 Effect of the Resizing Policy

Switching from letterboxing to square resizing (Figure 5) gains 4.8 balanced-accuracy points and reduces uncalibrated target ECE by 43% (0.0316 to 0.0181). No calibrator can change accuracy, and the ECE reduction is comparable to that of source-only temperature scaling under letterboxing (37% for C1 to 47% for C3). The mechanism follows from the audit: Kermany images have a median size of 512×496 pixels (aspect ratio 1.03) and OCTDL images 882×321 (aspect ratio 2.74). Letterboxing preserves this difference and places target images in a thin band between black bars, unlike anything seen in training; square resizing stretches both cohorts to fill the frame and removes the dominant geometric difference. This is a preprocessing choice that normalises the difference dominating this cohort pair, not a general domain-adaptation method, and it would not be expected to help where cohorts share an aspect-ratio distribution.

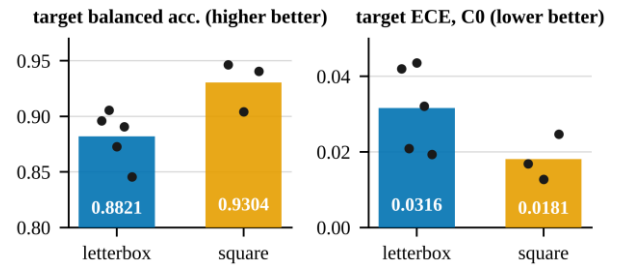


Figure 5. Effect of the resizing policy on the unseen scanner (ResNet-18, uncalibrated). Bars give the mean; dots are individual runs.

6.3 What the Scanner Change Does (H1)

Hypothesis H1 states that reliability degrades faster than discrimination. Figure 6 reports the error ratio $R(m)$ of Equation 7 for every metric. Under letterboxing, macro-AUROC error degrades 7.02× and 7.77× against ECE’s 4.96× and 3.36×, so discrimination degrades more and H1 is rejected. Under square resizing the ordering reverses for ECE (2.74×) and classwise ECE (2.96×) against AUROC error (2.18×), so H1 is supported for these two metrics, but not for Brier score (1.61×) or NLL (1.65×). A study reporting a single preprocessing policy would therefore have reached a confident verdict in either direction. Whether calibration is disproportionately fragile depends on how much geometric domain difference the preprocessing leaves in, and is not an intrinsic property of the modality.

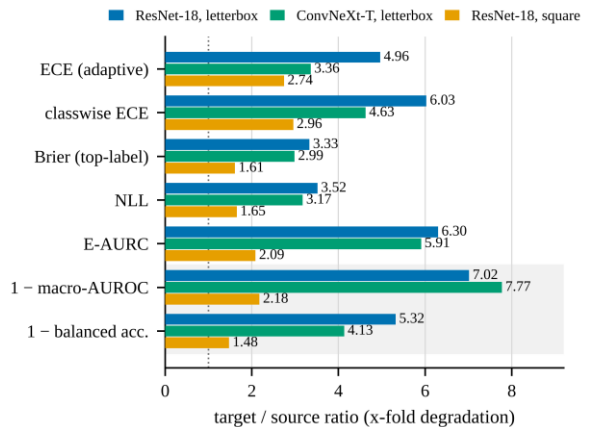


Figure 6. Source-to-target degradation ratio $R(m)$ (mean over runs). Shaded rows: discrimination metrics expressed as error. A ratio of 1 (dotted) means no degradation.

6.4 Calibration Conditions on the Unseen Scanner

Table 2. Target ECE (adaptive), mean $\pm$ SD (population) across runs

Condition	R18-LB	CNX-LB	R18-SQ
C0 none	0.0316 $\pm$ 0.0102	0.0298 $\pm$ 0.0047	0.0181 $\pm$ 0.0050
C1 TS clean	0.0199 $\pm$ 0.0073	0.0149 $\pm$ 0.0024	0.0166 $\pm$ 0.0062
C2 TS+generic	0.0183 $\pm$ 0.0055	0.0129 $\pm$ 0.0022	0.0170 $\pm$ 0.0058
C3 TS+physical	0.0166 $\pm$ 0.0057	0.0122 $\pm$ 0.0026	0.0249 $\pm$ 0.0113
C4 ICT+physical	0.0454 $\pm$ 0.0041	0.0181 $\pm$ 0.0050	0.0358 $\pm$ 0.0057
C4g ICT+generic	0.0450 $\pm$ 0.0034	0.0180 $\pm$ 0.0001	0.0412 $\pm$ 0.0077
C5 target ref.	0.0122 $\pm$ 0.0025	0.0096 $\pm$ 0.0009	0.0161 $\pm$ 0.0031

Table 2 and Figure 7 give target ECE for every condition. Source-only temperature scaling (C1) lowers ECE in 10 of 11 runs (5/5, 3/3, 2/3) and, in the letterboxed runs, closes 60% (R18-LB) and 74% (CNX-LB) of the gap to the target-fitted reference without target data. Under square resizing little gap remains (C0 0.0181, C1 0.0166, reference 0.0161), so calibration repair and preprocessing act as substitutes. Adding simulated shift to the calibration set beats plain temperature scaling in 8 of 11 runs (generic) and 7 of 11 (physical); in the letterboxed configurations the mean ECE reduction relative to C1 is 0.0016 and 0.0020 (generic) and 0.0033 and 0.0027 (physical).

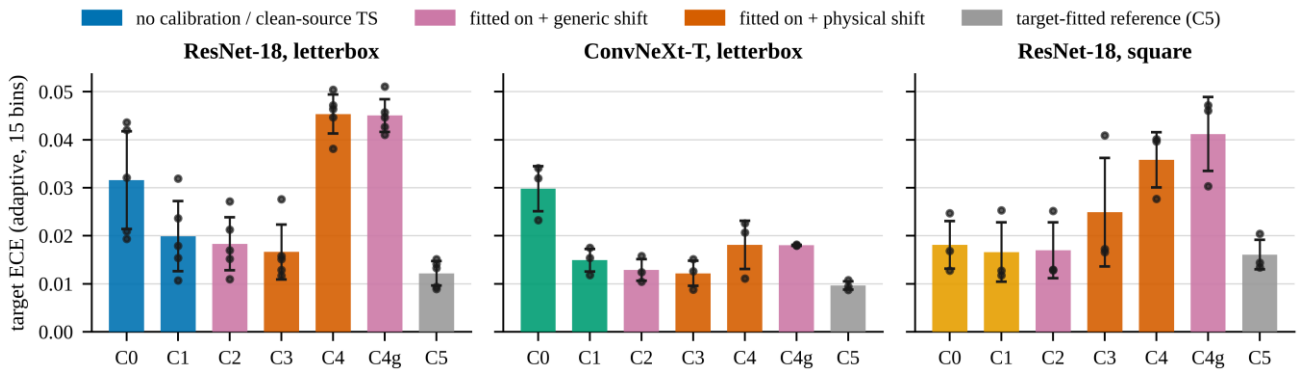


Figure 7. Target ECE by calibration condition for all three configurations (bars: mean; error bars: population SD across runs; dots: individual runs).

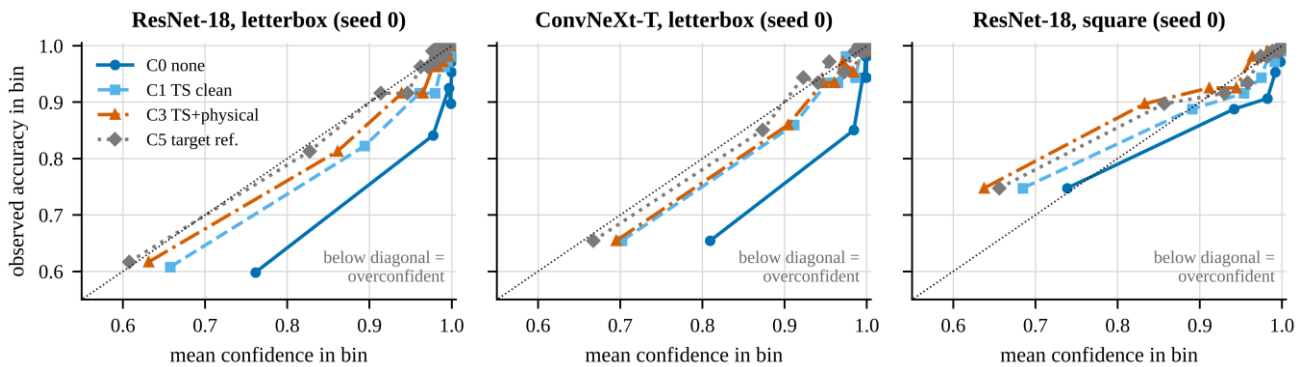


Figure 8. Reliability diagrams on the unseen scanner (seed 0 of each configuration; 15 equal-mass bins). Points below the dotted diagonal indicate over-confidence.

Within-run bootstrap intervals show how much of this is resolvable in a single model. The 95% interval of C0 target ECE spans 0.017 to 0.027 in width across the eleven runs, comparable to many differences between conditions. In paired bootstrap comparisons, C1 is better than C0 with an interval excluding zero in 6 of 11 runs (3/5, 3/3, 0/3), and C4 is worse than C3 in 7 of 11 runs; C3 versus C2 is discussed in Section 6.5.

The fitted temperatures explain these results. All source-fitted temperatures exceed 1, confirming that the networks are over-confident. Under letterboxing the target-fitted temperature is larger than the clean-source one (mean C5 1.682 versus C1 1.360 for R18-LB; 1.726 versus 1.499 for CNX-LB), so C1 under-corrects, and adding degraded data pushes the temperature towards the target value (C3 1.573 and 1.574). Under square resizing the target needs almost no extra softening (C5 1.169 versus C1 1.117), and the physical family over-shoots (C3 1.339), which is why C3 is worse than C1 in

that configuration. Figure 8 shows the corresponding reliability diagrams: uncalibrated letterboxed models lie well below the diagonal (over-confident), and C1, C3 and C5 move successive bins towards it.

6.5 Physical Versus Generic Shift (H2)

Hypothesis H2 states that the physical family calibrates better than the generic family at matched severity. Figure 9 shows the per-run differences in target ECE (C3 – C2; negative favours the physical family). The mean difference is -0.00170 (SD 0.00226) for R18-LB, -0.00069 (SD 0.00073) for CNX-LB and $+0.00793$ (SD 0.00549) for R18-SQ, where all three runs are positive ($+0.00445$, $+0.00365$, $+0.01569$). The physical family is better in 5 of 11 runs (3/5, 2/3, 0/3): under letterboxing its advantage is smaller than its across-seed SD and the sign flips, and under square resizing it is worse in every run and worse than plain C1 in 3 of 3 runs. In paired bootstrap tests no run shows the physical family significantly better, while two

runs (R18-LB seed 2, R18-SQ seed 2) show it significantly worse. H2 is therefore not supported in any configuration. This is a finding about this physical family on this cohort pair, not a claim that OCT physics is irrelevant to calibration. A plausible reason for the square-resizing result is that stretching rescales the two image axes by different per-image factors, so an “axial” degradation no longer has a consistent physical direction.

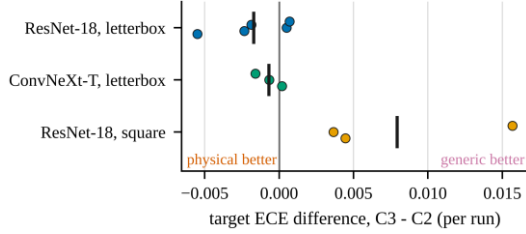


Figure 9. Per-run difference in target ECE between physical (C3) and generic (C2) calibration families. Negative values favour the physical family; vertical bars mark the mean

6.6 Input-Conditioned Temperature

Table 3. Target NLL and macro-AUROC for selected conditions (mean over runs)

Condition	R18-LB	CNX-LB	R18-SQ
NLL, C0 none	0.2324	0.1928	0.1687
NLL, C1 TS clean	0.1942	0.1541	0.1632
NLL, C4 input-cond.	0.7906	0.1779	0.5340
NLL, C5 reference	0.1860	0.1512	0.1607
AUROC, C1 TS clean	0.9845	0.9860	0.9873
AUROC, C4 input-cond.	0.9658	0.9859	0.9780

C4 is worse than scalar temperature scaling in 10 of 11 runs (5/5, 2/3, 3/3), and C4g in 11 of 11; the stronger claim that it is worse than no calibration holds in only 7 of 11 runs and is not made. Table 3 shows the failure in other metrics: on R18-LB the target NLL of C4 is 0.7906 against 0.1942 for C1, and because per-image temperatures re-order confidences across images, macro-AUROC falls from 0.9845 to 0.9658. The temperature network is fitted on source features and queried on shifted target features, i.e. off-distribution; a single global scalar has no such failure mode.

6.7 Patient-Level Aggregation (H3)

Hypothesis H3 states that patient-level reliability is not predictable from B-scan-level reliability. For the 618 patients represented in the de-duplicated target set (Section 3.1), probabilities are aggregated by mean probability, mean logit or the most confident scan. Balanced accuracy barely changes, but ECE varies across the four levels by factors (largest over smallest) of 1.70 (R18-LB; max-confidence 0.0537 versus B-scan 0.0316), 1.19 (CNX-LB) and 2.36 (R18-SQ; mean probability 0.0427 versus B-scan 0.0181). H3 is supported, but the worst rule changes with configuration, so the rule should be measured rather than fixed in advance.

6.8 Transferred Operating Points

Figure 10 translates calibration into referral behaviour. A threshold chosen on the source calibration split to promise 0.5% selective error delivers 0.34% to 0.68% on the source test split but 1.59% (R18-LB), 1.25% (CNX-LB) and 1.19% (R18-SQ) on the target, 2.4 to 3.2 times the promised risk, while retaining 77.3%, 81.3% and 69.7% of target cases; at a promised 1% the target delivers 2.67%, 2.63% and 1.87%. Re-

deriving the threshold after C1 or C3 leaves these values essentially unchanged (1.60% and 1.59% for R18-LB at 0.5%), because scalar temperature scaling preserves the ranking of cases.

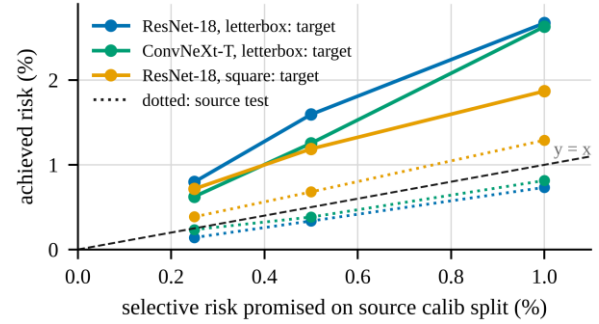


Figure 10. Promised versus achieved selective risk when a referral threshold fixed on source data is transferred unchanged (uncalibrated model). Solid: unseen scanner; dotted: source test split

7. DISCUSSION

The most informative results did not come out as hypothesised: reliability is not stably the fragile axis, and physical faithfulness of the simulated shift did not reliably help within eleven runs. The practical prescription is to fix the resizing policy first, use plain source-fitted temperature scaling by default, prefer the least flexible calibrator under device shift, choose the patient aggregation rule by measurement, and collect a small labelled target sample before trusting a transferred referral threshold.

Three single-configuration conclusions (input-conditioned scaling worse than no calibration, a 70% ECE inflation from patient aggregation, and disproportionately fragile reliability) did not survive replication across configurations.

8. LIMITATIONS

There is one source–target pair, with manufacturer, site and population shifting together, and three configurations with three to five seeds each. The physical family rests on an assumed 45 dB log-compression inverse, and per-image standardisation nulls global tone-mapping axes for both families. All experiments resize to 224×224; a native-resolution study might reveal a physical-versus-generic difference and avoid the axis-rescaling problem of Section 6.5. OCTDL classes without a source counterpart (354 images) are excluded, the target DME class is small (139 images), and 177 contradictory-label source images were retained. Some run counts depend on the metric: with equal-width ECE the physical family is better in all three CNX-LB runs (mean −0.00072), with classwise ECE it is worse in all five R18-LB runs, and under square resizing C1 improves classwise ECE and Brier score in only 1 of 3 runs. No claim of clinical readiness is made.

9. CONCLUSION

Under a change of OCT manufacturer, adaptive ECE rose 2.7- to 5.0-fold; discrimination error rose by a larger factor under letterboxing and a smaller one under square resizing. Source-fitted temperature scaling recovered 60% to 74% of the deployable calibration gap in letterboxed runs and lowered ECE in 10 of 11 runs; a physically faithful simulated shift did not reliably beat generic corruptions (5 of 11 runs); input-conditioned temperatures lost to a global temperature in 10 of 11 runs; and transferred referral thresholds delivered 2.4 to 3.2 times their promised error. Square resizing improved accuracy, which no calibrator can, and reduced ECE about as much as the

best source-only calibrator, and the two leakage channels found in Kermany and OCTDL are not closed by per-subject splitting alone.

10. REFERENCES

- [1] Vijayan, M., Veerapu, G., and Nisha, J. S. 2026. A spatial adaptive multi-scale ConvNeXt framework for robust, calibrated, and statistically validated multi-class retinal OCT classification. *Frontiers in Artificial Intelligence*, 9. DOI: 10.3389/frai.2026.1898978.
- [2] Kumar, A. 2026. Calibrated hybrid CNN-Transformer for retinal OCT classification. *arXiv:2607.09809*.
- [3] Chutani, D. and Yalavarthy, P. K. 2026. Noise-aware dynamic convolution for improved generalizability of retinal disease diagnosis using optical coherence tomography images. *Journal of Biomedical Optics*, 31(8), 086006. DOI: 10.1117/1.JBO.31.8.086006.
- [4] Lopukhova, E. A., Idrisova, G. M., Mukhamadeev, T. R., Voronkov, G. S., Kutluyarov, R. V., and Topolskaya, E. P. 2026. A hierarchical deep learning architecture for diagnosing retinal diseases using cross-modal OCT to fundus translation in the lack of paired data. *Journal of Imaging*, 12(1), 36. DOI: 10.3390/jimaging12010036.
- [5] Morano, J., Fazekas, B., et al. 2025. Multimodal foundation model and benchmark for comprehensive retinal OCT image analysis. *npj Digital Medicine*, 8, 576. DOI: 10.1038/s41746-025-01852-3.
- [6] Salvador, T., Voleti, V., Iannantuono, A., and Oberman, A. 2021. Frustratingly easy uncertainty estimation for distribution shift. *arXiv:2106.03762*.
- [7] Zhang, Y., Xu, C., Wu, Y., Guan, Z., and Zhao, W. 2025. Target-agnostic calibration under distribution shift with frequency-aware gradient rectification. *arXiv:2508.19830*.
- [8] Hendrycks, D. and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. ICLR*. *arXiv:1903.12261*.
- [9] Agnihotri, S., Schader, D., Sharei, N., Kacar, M. E., and Keuper, M. 2025. Are synthetic corruptions a reliable proxy for real-world corruptions? In *CVPR 2025 Workshop on Synthetic Data for Computer Vision*. *arXiv:2505.04835*.
- [10] Tampu, I. E., Eklund, A., and Haj-Hosseini, N. 2022. Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Scientific Data*, 9, 580. DOI: 10.1038/s41597-022-01618-6.
- [11] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *Proc. ICML*. *arXiv:1706.04599*.
- [12] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Proc. NeurIPS*. *arXiv:1906.02530*.
- [13] Wang, X., Long, M., Wang, J., and Jordan, M. I. 2020. Transferable calibration with lower bias and variance in domain adaptation. In *Proc. NeurIPS*. *arXiv:2007.08259*.
- [14] Ding, Z., Han, X., Liu, P., and Niethammer, M. 2021. Local temperature scaling for probability calibration. In *Proc. ICCV*. *arXiv:2008.05105*.
- [15] Joy, T., Pinto, F., Lim, S.-N., Torr, P. H. S., and Dokania, P. K. 2023. Sample-dependent adaptive temperature scaling for improved calibration. In *Proc. AAAI*. *arXiv:2207.06211*.
- [16] Khan, B. and Syed, T. Q. 2025. Confidence-calibrating regularization for robust brain MRI segmentation under domain shift. *arXiv:2509.23176*.
- [17] Di Salvo, F., Doerrich, S., and Ledig, C. 2024. MedMNIST-C: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions. *arXiv:2406.17536*.
- [18] Kermany, D. S., Goldbaum, M., Cai, W., et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5), 1122–1131. DOI: 10.1016/j.cell.2018.02.010.
- [19] Kulyabin, M., Zhdanov, A., Nikiforova, A., et al. 2024. OCTDL: Optical coherence tomography dataset for image-based deep learning methods. *Scientific Data*, 11, 365. DOI: 10.1038/s41597-024-03182-7.