IMAGE-DEPENDENT CALIBRATION MAPS FOR COMPOUND SEGMENTATION LOSSES

Anonymous authors
Paper under double-blind review

ABSTRACT

In many domains, a segmentation network has to provide more than a binary decision: a calibrated probability map. For example, in radiotherapy planning, treatment margins and dose can be defined based on the probability that a specific area (i.e., voxel in the image) belongs to a tumor. However, models trained with compound losses, such as cross-entropy plus soft Dice, encode different posterior probabilities into the same output. This shift is caused by the optimum forced by the compound loss. By balancing the opposing loss gradients at a stationary point, we derive a calibration map $\Psi_{f,D}$ in closed form, parameterized by two properties of the image, the foreground fraction f and the attained Dice D. We describe this map for the example of the cross-entropy plus soft Dice compound loss, and show that no beta, Platt, or temperature map can represent $\Psi_{f,D}$, and no map fitted once per model can recover the true posterior of multiple images. In addition to $\Psi_{f,D}$, we propose a foreground-conditioned version of isotonic regression that has sufficient capacity to model $\Psi_{f,D}$ and improves on standard isotonic regression. Without any held-out labels or retraining, correcting with $\Psi_{f,D}$ alone consistently reduces NLL and ECE scores compared to raw outputs. Furthermore, applying a beta or isotonic map fitted on held-out data in conjunction with $\Psi_{f,D}$ is more effective than the respective fitted map alone. Both $\Psi_{f,D}$ and the foreground-conditioned isotonic regression improve the placement of uncertainty margins across three tumor segmentation tasks, measured by the sample-wise IoU of the 0.25–0.75 probability margin.

1 INTRODUCTION

Segmentation models are typically supervised with binary masks because they are easier to generate. Yet many downstream decisions, especially in high-stakes domains such as medicine, depend on the continuous confidence values the networks produce. In tumor segmentation, for example, the relevant quantity is not only whether a voxel is classified as tumor, but the probability that it belongs to the tumor region. Treatment margins and dose, e.g., in radiotherapy planning, can then be defined relative to a chosen probability level (Samarasinghe et al., 2021). The same applies whenever a segmentation informs decisions that depend on uncertainty, such as monitoring lesion growth or determining whether a structure should be treated.

Despite this importance of calibrated network outputs, the reading as true posterior probabilities is often unavailable in practice. One important reason is the *generalization gap*: even models trained with a proper scoring rule, such as CE, whose optimum is the true posterior, are miscalibrated on unseen data (Guo et al., 2017), and this miscalibration is often estimated and corrected via a single calibration map obtained from held-out data (e.g., Guo et al., 2017; Zadrozny & Elkan, 2002).

However, models trained with compound segmentation losses are miscalibrated beyond the generalization gap because the optimum forced by such a compound loss does not sit at the posterior probability, even for a perfectly fitted model without generalization gap. Compound losses are standard in the field and used, e.g., by nnU-Net (Isensee et al., 2021), the dominating framework for medical image segmentation, which combines cross-entropy (CE) plus soft dice as a loss function. We call this effect *loss-induced* miscalibration. The outputs are shifted, and the shift is a deterministic function of the loss, obtained by balancing its two opposing gradients at the optimum.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we derive a function that expresses the loss-induced miscalibration for the Dice+CE loss parameterized by two properties of the image. (1.) The foreground fraction f, defined by the predicted and ground truth foreground, and (2.) the Dice D, attained on the image. Figure 1 shows the family this produces. The image dependence of the shift is why a map fitted once per model cannot recover the per-image posterior. For example, at the same fit quality $D = 0.9$, an output of 0.9 carries a posterior of 0.69 when a tenth of the image is foreground, and 0.50 when a hundredth is. Consequently, the contour drawn at a chosen threshold is not the region whose posterior exceeds that threshold, and how far it misses depends on the image.

Furthermore, we propose a calibration map in closed form for CE + Dice (Sec. 3.3), and propose how to estimate (f, D) from the model outputs alone (Sec. 4). Moreover, we define what a replacement of the map fitted on held-out data requires in order to replace the derived one (Sec. 5): (1.) The capacity to express the shape, and (2.) a parameterization that varies per image. Based on these insights, we develop an image-conditioned improvement for isotonic regression and beta calibration.

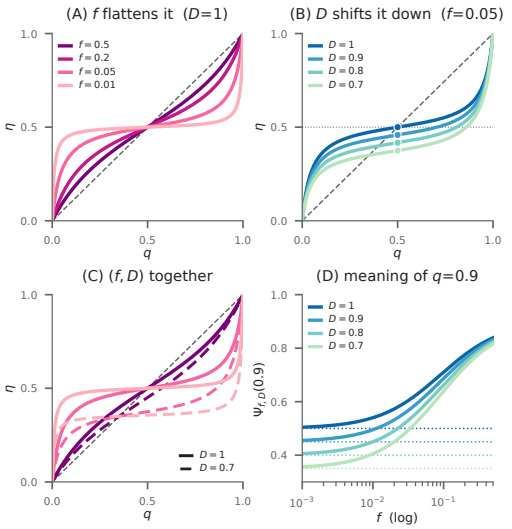

Figure 1: **Post-hoc calibration maps for compound losses are image-dependent.** We derive a closed-form solution (6) for calibrating the Dice+CE loss, parameterized by the foreground fraction f and soft-Dice D. **(A)** f flattens the map. **(B)** D shifts the map downwards. **(C)** Examples for different (f, D) pairs. **(D)** The map's value $\Psi_{f,D}(0.9)$ against f.

We evaluate the real-world applicability of the derived calibration map using NLL and expected calibration error across five datasets. Beyond this, we evaluate the post hoc calibration method to identify the label-derived 0.25–0.75 posterior margin across three lesion datasets.

2 Related Work

The poor calibration of Dice-trained models has been established in prior work. Mehrtash et al. (2020) report substantial miscalibration relative to cross-entropy, Bertels et al. (2021) trace a systematic volume bias, and Popordanoska et al. (2021) connect that bias to calibration. Nordström et al. characterize the optima of Accuracy and Dice under label noise, show soft Dice's minimizers lie at the boundary of $[0, 1]$ except on a single level set, adapt a calibrated surrogate to Dice, and give a thresholding rule recovering Dice-optimal decisions from a cross-entropy model (Nordström et al., 2020; 2022; 2023a;b). Lux et al. (2026) propose a gradient-surgery for calibrating region-based losses, Yeung et al. (2023) adapt the Dice loss to improve calibration and Barfoot et al. (2024) propose a loss to directly optimize for ECE.

Temperature scaling (Guo et al., 2017), isotonic regression (Zadrozny & Elkan, 2002) and their relatives are the standard post hoc remedy and transfer to segmentation, with mixed results. The two have been compared since Niculescu-Mizil & Caruana (2005). Rousseau et al. (2025) find isotonic regression and beta calibration effective and temperature scaling the weakest parametric method, and, because dataset-level calibration can hide per-subject miscalibration (Jungo et al., 2020), also score each image, where a fitted map improves most images but not all. Ding et al. (2021) go beyond one map per dataset and use a network trained on a held-out split to predict a temperature per image and, in their local variant, per voxel.

Decomposing a loss into a proper loss and an invertible link is due to Reid & Williamson (2010). Charoenphakdee et al. (2021) recover the posterior from a focal minimizer, Komisarenko & Kull (2024) show two temperature scalings bracket that transformation, and Bao & Sugiyama (2020) take the framework towards calibrated surrogates for linear-fractional utilities. All of it treats *pointwise* losses, whereas properties of the image define the link of the compound losses used in practice.

We correct the probability shift induced by the compound loss using an image-specific map, without retraining or held-out labels. This allows identical network scores to correspond to different posterior probabilities across images.

3 WHERE A COMPOUND LOSS SHIFTS THE OUTPUTS

This section derives the closed-form function for the relation between the network's output and the true posterior probability given the foreground fraction f and the attained Dice score D. Proofs are in Sup. K. We use the following notation. Given an image patch x, the label Y_i at voxel i is a random variable, with $\eta_i = \mathbb{P}(Y_i{=}1 \mid x)$ the posterior probability that i is foreground; the network outputs $q_i \in (0, 1)$ at every voxel as a function of the image alone; and training minimizes the conditional risk $R(q) = \mathbb{E}[L(q, Y) \mid x]$.

3.1 THE BALANCE CONDITION

We decompose a binary loss $\ell(q)$ into its two branches $\ell_1(q)$, where the label is 1, and $\ell_0(q)$, where the label is 0, and write their derivatives with the sign of the direction in which a gradient step moves q,

$$\underbrace{g_1(q) = -\ell_1'(q)}_{\text{how hard the loss pulls } q \textit{ up}} \quad , \qquad \underbrace{g_0(q) = \ \ \ell_0'(q)}_{\text{how hard it pulls } q \textit{ down}} . \tag{1}$$

A gradient step on a voxel with label 1 raises q in proportion to g_1, and on a voxel with label 0 lowers it in proportion to g_0, so a voxel with posterior η is subject to the upward pull a fraction η of the time and the downward pull otherwise.

Proposition 1 (Pulls balance at the minimal risk)**.** *Let $R(q) = \eta\, \ell_1(q) + (1 - \eta)\, \ell_0(q)$ be the risk at a voxel with posterior η. If R is minimized at an interior q^*, then*

$$\underbrace{\eta\ g_1(q^*)}_{\textit{expected pull up}} \ = \ \underbrace{(1 - \eta)\ g_0(q^*)}_{\textit{expected pull down}}, \tag{2}$$

and therefore

$$\boxed{\eta \ = \ \Psi(q^*) \ = \ \frac{g_0(q^*)}{g_1(q^*) + g_0(q^*)}.} \tag{3}$$

We call $\Psi\colon q \mapsto \eta$ the *calibration map*; in the composite-loss view of Reid & Williamson (2010) it is the inverse link. For a pointwise loss, it is the posterior-recovering transformation described in Charoenphakdee et al. (2021).

3.2 WHAT A REGION TERM ADDS

Compound losses add a region term (L^r) to a pointwise one (L^p) with the compound loss being $L = L^{\mathrm{p}} + L^{\mathrm{r}}$. Therefore, L's gradient at voxel j no longer depends only on j, and Proposition 1 does not apply directly. Let N be the number of voxels in the patch the loss is computed on, $P = \sum_i q_i$ the predicted foreground mass, and $S_Y = \sum_i Y_i$ and $I_Y = \sum_i q_i Y_i$ the ground truth masks volume and its intersection with the output. Soft Dice couples all voxels through these global sums. Differentiating soft dice at voxel j yields:

$$N\,\frac{\partial L^{\mathrm{r}}}{\partial q_j} = \begin{cases} -a(Y), & Y_j = 1, \\ +b(Y), & Y_j = 0, \end{cases} \qquad a(Y) = \frac{2 - D_Y}{2 f_Y}, \quad b(Y) = \frac{D_Y}{2 f_Y}, \tag{4}$$

where $f_Y = (P + S_Y)/2N$ is the mean of the predicted and the true foreground fraction, and $D_Y = 2I_Y/(P + S_Y)$ is the soft Dice attained against the mask. The $1/N$ matches CE's per-voxel mean and cancels from (3), so the patch size never enters the map. The mask is a random variable, so the two pulls are too. To obtain a calibration map for compound losses, we add these global terms to the pointwise gradients.

Proposition 2 (Region pulls)**.** *At a stationary point of R, every voxel j satisfies Eq.* (2) *with $g_1 = g_1^{\mathrm{p}} + a_j$ and $g_0 = g_0^{\mathrm{p}} + b_j$, where $a_j = \mathbb{E}[\, a(Y) \mid x,\, Y_j = 1\,]$ and $b_j = \mathbb{E}[\, b(Y) \mid x,\, Y_j = 0\,]$.*

The map we derive evaluates the two pulls at the mean sums instead, $S = \sum_i \eta_i$ and $I = \sum_i q_i \eta_i$,

$$g_1 = g_1^{\mathrm{p}} + a, \qquad g_0 = g_0^{\mathrm{p}} + b, \qquad a = \frac{2 - D}{2f}, \quad b = \frac{D}{2f}, \tag{5}$$

with $f = (P + S)/2N$ and $D = 2I/(P + S)$. Unless stated otherwise, f and D denote these two quantities of a patch: f is the mean of its predicted and its expected true foreground fraction, and D the soft Dice between the output and the posterior. The sum $a + b = 1/f$ does not depend on D, so the foreground fraction alone fixes the total pull of the region term, and halving the object doubles it. The attained Dice divides this pull between the two label branches, with $a = b$ if and only if $D = 1$, and an imperfect model has $a > b$: its foreground is pulled up harder than its background is pulled down.

3.3 The closed form for CE+Dice

Substituting CE's pulls $1/q$ and $1/(1 - q)$ plus the two constants of Eq. (5) into Eq. (3) yields a closed form.

Proposition 3 (Closed form)**.** *For* $L = \mathrm{CE} + \mathrm{Dice}$ *at a stationary point with the pulls Eq.* (5) *at foreground fraction* f *and attained Dice* D*,*

$$\boxed{\eta \;=\; \Psi_{f,D}(q) \;=\; \frac{q\left[\, f + \frac{1}{2} D\,(1 - q)\,\right]}{f + q(1 - q)}} \tag{6}$$

for every $q \in (0, 1)$.

The role of f *and* D*.* As the object grows, the region term weakens, and Ψ relaxes towards q without reaching it, since $f \leq 1$ keeps the total pull $1/f$ at least 1. As $f \to 0$, the map tends to $D/2$, a constant independent of the output. The map is symmetric only in the perfect Dice case: $\Psi_{f,D}(1 - q) = 1 - \Psi_{f,D}(q)$ for all q if and only if $D = 1$. A model achieving perfect Dice is overconfident symmetrically about $\frac{1}{2}$. Figure 1 shows examples of maps for different pairs of (f, D).

The optimal decision threshold moves. $\Psi_{f,D}(q^*) = \frac{1}{2}$ at the root $q^* \in (0, 1)$ of $(1 - D)\, q^2 + (2f - 1 + D)\, q - f = 0$, which equals $\frac{1}{2}$ if and only if $D = 1$, so thresholding at 0.5 does not threshold at $\eta = \frac{1}{2}$ unless the model is perfect. The discrepancy grows as the object shrinks, at $D = 0.9$ the output corresponding to $\eta = \frac{1}{2}$ is 0.71 when $f = 0.05$ and 0.91 when $f = 0.01$. Empirical work also finds 0.5 a poor threshold (Bice et al., 2021; Hussein et al., 2023) without computing a per-image value.

3.4 Assumptions

Remark 1 (Necessary assumptions). The approximation we make regarding the mean sums, $S = \sum_i \eta_i$ and $I = \sum_i q_i \eta_i$, is exact when the mask's volume does not vary given the image, and when the Dice is computed against the marginal labels η rather than against sampled masks (Nordström et al., 2023b). Otherwise, Eq. (5) is an approximation whose error grows with the spread of the volume. Supplement E measures the approximation with a controlled volume spread against a known posterior. The map still removes 85–89% of the shift when the volume's spread reaches 50%. Moreover, we assume that the outputs are at a stationary point of the conditional risk at every voxel. We do not expect these assumptions to hold perfectly in real-world applications, but we find empirically that we are close enough to the idealized setting for the assumptions to remain useful. Section 4 estimates (f, D) for a patch at inference time.

4 Estimating f and D at inference time

Section 3 derives the shift that is loss optimal and that the shape of the shift can be parameterized by f and D. This section tries to answer if networks can adapt their outputs depending on f and D of a patch at hand. At inference time, no labels exist, hence S and I can only be *estimated* by a model. A shift of the outputs is therefore governed by the f and D that a model expects for an image at hand.

Can models estimate f and D? A network that cannot estimate the foreground fraction f and the fit quality D of its own prediction produces the same output for every voxel of posterior η. It settles where the pulls balance across the entire training patch stream and cannot adapt in an image-dependent manner. For the fixed-size patches for nnU-Net, that stream acts as one *concatenated* image, and the loss-optimal pair is its pooled statistics,

$$f^\star = \frac{1}{|K|}\sum_{k\in K}\frac{P_k + S_k}{2N}, \qquad D^\star = \frac{2\sum_{k\in K} I_k}{\sum_{k\in K}(P_k + S_k)}, \qquad K = \{k : S_k > 0\}, \tag{7}$$

with P_k and I_k measured from the model's outputs. Sup. §B derives (7).

For networks to shift their output based on the per-image f and D, they need access to global statistics. This access is given for most commonly used networks, e.g., via a global receptive field (in deep U-Nets) or via instance norm (in nnUNet). We empirically show that common network architectures are capable of the required estimation in Sup. D.2.

At inference time, the labels needed to compute S and I are unavailable. However, their expectation is known by Prop. 3: $\mathbb{E}[Y_i \mid x] = \Psi_{f,D}(q_i)$, i.e., exactly the posterior returned by the map. Replacing $Y_i \mapsto \Psi_{f,D}(q_i)$ and $q_i Y_i \mapsto q_i \Psi_{f,D}(q_i)$ closes the two definitions of Eq. (5) into a pair in (f, D) alone,

$$f = \frac{1}{2N}\sum_i \big[\, q_i + \Psi_{f,D}(q_i)\,\big], \qquad D = \frac{2\sum_i q_i\,\Psi_{f,D}(q_i)}{\sum_i q_i + \sum_i \Psi_{f,D}(q_i)}, \tag{8}$$

whose fixed point is the pair consistent with the outputs under Rem. 1. A network that sees only the image can depend on the label sums S and I only through their expectation given the image, which Eq. (8) uses in place of the labels. On held-out labels, its f is within 6% and its D within 0.06 of the label-measured pair on nine of ten models (Sup. Table 3). We solve for f and D using a root-finder in the box $f \in (0, 0.99)$, $D \in (10^{-3}, 1)$ started at $(0.1, 0.75)$.

Batch pooled Dice loss. Where nnU-Net's planner sets `batch_dice`, the Dice loss is computed over the B patches of a batch as a single concatenated image, of fraction $f_{\text{avg}} = \frac{1}{B}(f + \Sigma_f)$, with f the image's own value from Eq. (8) and Σ_f the sum over its $B-1$ training partners. The pull scales as $1/f_{\text{avg}}$, and averaging it over the partners gives

$$f_{\text{eff}} \;=\; 1/\mathbb{E}[\,1/f_{\text{avg}}\,], \qquad D_{\text{eff}} \;=\; f_{\text{eff}}\;\mathbb{E}[\,D_{\text{batch}}/f_{\text{avg}}\,], \tag{9}$$

with D_{batch} the Dice of the concatenated batch. The image's own pair thus enters at a weight of $1/B$, so pooling dilutes the image dependence (Sup. C). Strictly, $\Psi(f_{\text{eff}}, D_{\text{eff}})$ should stand in for the labels in Eq. (8), one coupled fixed point; it moves f_{eff} by a median of under 2% and D_{eff} by 0.01–0.04 (Sup. C).

The procedure, in short. Once per model trained with a pooled Dice, collect the per-patch f_k, P_k and I_k of the training patches the sampler draws. Per test image, solve Eq. (8) for the image's own (f, D), blend it with the $B-1$ training partners by Eq. (9), which at $B = 1$ returns it unchanged, and map every voxel through $\Psi(f_{\text{eff}}, D_{\text{eff}})$ of Eq. (6).

5 Fitted calibration maps

In practice, post hoc calibration is used even for models trained with proper scoring rules, such as CE, to compensate the miscalibration from imperfect generalization. This motivates fitting both sources of miscalibration jointly where enough held-out labeled data is available. A fitted map that can replace Ψ must meet two requirements. (1.) The capacity to express the shape of Ψ, and (2.) a parameterization that varies per image.

Capacity. Beta calibration (Kull et al., 2017; 2019), the standard parametric family, which contains Platt and temperature scaling as special cases, can only move steepness from one end of the range to the other, whereas $\Psi_{f,D}$ is steeper than the identity at both ends: $\Psi(q)/q \to \lambda_0 = 1 + D/2f$ as $q \to 0$ and $(1 - \Psi(q))/(1 - q) \to \lambda_1 = 1 + (2 - D)/2f$ as $q \to 1$, so $\lambda_0\lambda_1 > 1$.

Proposition 4 (Shape)**.** *A beta map,* $\operatorname{logit}\mu = c + \alpha \log q - \beta \log(1 - q)$*, has finite non-zero end slopes only when* $\alpha = \beta = 1$*. It is then the shift* $\sigma(\operatorname{logit} q + c)$*, whose end slopes multiply to exactly* 1*. Hence no beta map, and therefore no Platt map and no temperature, equals* $\Psi_{f,D}$ *at any* (f, D)*.*

Temperature and Platt scaling also miss the asymmetry of Ψ at $D < 1$ (Sec. 3.3), so their error grows as D falls, while beta absorbs the asymmetry but not the shape (Figure 2).

Image dependence. By Eq. (6), the maps of two images with different f and the same D differ at every output except 0, 1, and the fixed point $D/2$ (Figure 1.D).

Proposition 5 (Image dependence)**.** *Let m be any map of q alone. On one of two images with pairs (f, D) and (f', D'), $|m(q) - \eta| \geq \frac{1}{2}\,|\Psi_{f,D}(q) - \Psi_{f',D'}(q)|$ at output q.*

Isotonic regression (Ayer et al., 1955; Zadrozny & Elkan, 2002), the nonparametric fit over monotone maps, can represent $\Psi_{f,D}$ at any fixed (f, D), since the map is monotone in q, but by Proposition 5 a single map fitted on a population of images cannot recover the per-image posterior.

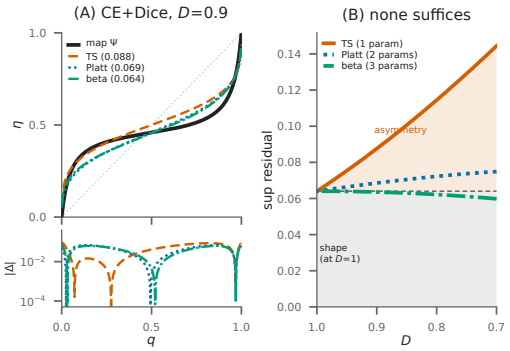

Figure 2: **No member of the beta family represents the map. (A)** best-case fit per family to CE + Dice at $D = 0.9$, $f = 0.05$, with the residual beneath. **(B)** the residual each family attains against D. Temperature and Platt scaling pay for the asymmetry as D falls; the full beta family absorbs it and stays at the shape floor of $D = 1$.

Image-conditioned fitted maps. To account for the image dependence, we propose to condition isotonic regression and beta calibration on the f that Ψ uses, estimated from the outputs by Eq. (8) and blended by Eq. (9) if the Dice is pooled. For a test image with estimate f, every training voxel is weighted by the distance of its patch's f_k in $\log f$, $w_k = \exp\!\big(-(\log f_k - \log f)^2/2h^2\big)$, and each family is fitted as before on the weighted voxels and applied to that image only. For $h \to \infty$, both reduce to one map per model, for $h \to 0$ to one map per image. We fix $h = 1$, a factor e in f, for every model (Sup. F). Both variants need labeled cases across the range of f.

6 Experiments

First, we evaluate the calibration results through the following questions. (1.) Does Ψ improve real model outputs without held-out labels? (2.) Do image-dependent f and D estimates improve compared to fixed values? (3.) How can both sources of miscalibration be corrected? (4.) Does conditioning the fitted maps on f improve compared to global maps?

A note on the evaluation. ECE is a population statistic, so a map matching $\mathbb{E}[Y \mid q]$ over all cases scores zero on it without being right on any image (Jungo et al., 2020), whereas Ψ targets the posterior of each image. NLL is a proper score averaged per case, so Table 1's NLL is per image and cannot collapse to zero for $\mathbb{E}[Y \mid q]$.

Second, we test the utility of Ψ and the proposed conditioned isotonic regression in identifying the 0.25-0.75 posterior probability margin for lesion datasets, estimated from the ground-truth labels.

6.1 Setup

We use nnU-Net v2 (Isensee et al., 2021) unmodified on fold 0 of five tasks, spleen, pancreas, colon, and liver tumor from the Medical Segmentation Decathlon (Antonelli et al., 2022), and the kidney tumor of KiTS23 (Heller et al., 2023), in both planned 3D configurations, each with a single foreground channel. The loss is the default `DC_and_CE_loss` with `do_bg=False`, the CE + soft-Dice objective of Sec. 3.3, under deep supervision, whose auxiliary heads leave Eq. (6) at the output head unchanged (Sup. F). The planner sets `batch_dice` to `False` for `3d_lowres` and `True` for `3d_fullres`, so the region pull is the patch's own in the first and the batch's in the second. We report ECE (15 uniform bins) and NLL on an evaluation region around the model's predicted foreground, so that the confidently empty background does not dominate the results (Sup. F.1). Baselines

	Panc. l		Panc. f		Spleen l		Spleen f		KiTS l		KiTS f		Colon l		Colon f		Liver l		Liver f	
method	ECE	NLL	ECE	NLL	ECE	NLL	ECE	NLL	ECE	NLL	ECE	NLL	ECE	NLL	ECE	NLL	ECE	NLL	ECE	NLL
no held-out labels: nothing is fitted (bold: best in this block)																				
raw output	.128	.866	.161	.902	.030	.163	.059	.242	.105	.718	.173	.872	.170	1.373	.177	1.175	.116	1.149	.242	1.369
$\Psi(f^\star, D^\star)$	.074	.443	.099	.489	**.017**	.110	.042	.184	.077	.406	.126	.599	.112	.670	.120	.589	.086	**.742**	.208	.897
$\Psi(f_{\text{eff}}, D^\star)$	**.068**	**.425**	**.070**	**.434**	.018	.105	**.035**	**.172**	.045	.306	**.072**	**.422**	**.090**	**.612**	**.096**	.482	.082	.752	.151	.693
$\Psi(f_{\text{eff}}, D_{\text{eff}})$	**.068**	.426	.072	.435	.018	**.104**	.047	.194	**.023**	**.289**	.074	.435	**.090**	.622	.097	**.479**	**.079**	.753	**.110**	**.666**
fitted on held-out labels, one map per model (bold: best in this block)																				
Ψ, (f, D) fitted	.053	.376	.036	.402	.013	.111	**.037**	**.175**	.057	.348	**.065**	.476	.075	.458	.093	.487	.084	.717	.153	.674
temperature	.066	.357	.053	.414	.017	.116	.040	.180	.038	.302	.086	.463	.058	**.421**	**.055**	**.433**	.073	.441	.134	**.574**
Platt	.063	.357	.053	.417	.014	.117	.038	.184	.038	.302	.082	.462	.043	.428	.110	.439	.075	**.438**	.140	.595
beta	.055	.356	.059	.418	.011	.113	.046	.195	.035	.301	.083	.464	**.042**	.431	.111	.486	.074	.445	**.124**	.585
isotonic	**.026**	**.334**	**.030**	**.400**	**.005**	**.107**	**.037**	.184	**.033**	**.300**	.069	**.460**	.055	.427	.108	.542	**.063**	.450	.130	.664
a fitted map chained after Ψ (bold: beats the same family unchained)																				
beta $\circ\, \Psi(f_{\text{eff}}, D_{\text{eff}})$	**.036***	**.338***	**.029**	**.396***	**.004**	**.097***	.048	.200	**.016**	**.252***	**.041**	**.416**	.052	**.422**	**.087**	**.452**	**.062**	.485	**.091**	**.560**
isotonic $\circ\, \Psi(f_{\text{eff}}, D_{\text{eff}})$	.027	**.333**	**.024**	**.389**	**.004**	**.095**	.064	.213	**.023**	**.255***	**.047**	**.414**	**.046**	**.417**	**.093**	**.443**	**.050**	**.408**	**.098**	**.603**
fitted on held-out labels, conditioned on the image (bold: beats the same family unconditioned)																				
IBTS (Ding et al., 2021)	**.047***	.358	.057	.415	—	—	—	—	.044	**.289**	**.065**	**.436**	—	—	—	—	—	—	—	—
beta$_f$	.055	.356	**.056**	**.416**	.012	**.112**	**.043**	**.191**	**.028**	**.273***	**.058**	**.419**	**.033**	**.429**	**.085**	**.476**	**.073**	**.379**	.140	.617
isotonic$_f$	.027	.334	**.028**	**.397***	.006	**.105**	.041	**.183**	**.015**	**.263***	**.044**	**.405**	**.047**	**.422**	**.084**	**.532**	.071	**.432**	**.123**	.666
hard Dice at 0.5	.828		.796		.970		.928		.837		.783		.755		.658		.725		.667	
sd log f (sampler)	.48		.51		.44		.45		1.76		1.56		1.10		1.03		2.02		2.16	
cases improved, ECE_{case}	79%		80%		89%		78%		67%		75%		78%		75%		76%		100%	
cases improved, NLL	98%		98%		100%		89%		95%		87%		100%		92%		100%		100%	

Table 1: Main results. We evaluate our proposed calibration map Ψ without fitting on held-out data (block 1), compare it with approaches fitted on held-out data (block 2), combine it with them (block 3), and condition fitted maps on the foreground fraction f (block 4). l/f: nnU-Net's low- and full-resolution 3D configurations. IBTS needs a held-out split large enough to train its network. Marks have different meanings in the different blocks (stated in the block header). Underline is the column's best overall result. In the last two blocks, a star marks $p < .01$ for that block's own comparison over cases, by an exact sign-flip permutation for NLL and a paired bootstrap for ECE. Supplement F provides further experimental details. The last four rows describe: the model's hard Dice at 0.5 on the scored patches, $\text{sd} \log f$ over the patches nnU-Net's sampler draws (Sup. Table 4), and the fraction of cases on which $\Psi(f_{\text{eff}}, D_{\text{eff}})$ lowers ECE and NLL against the raw output, with ECE computed inside each case (Sup. F).

are temperature, Platt, beta, and isotonic fitted leave-one-case-out on the evaluation labels, and the image-based temperature scaling (IBTS) of Ding et al. (2021). For this method, a 3D network reads the logit map and the image and predicts one temperature per patch. The prediction models are trained using NLL on the other cases, following the same leave-one-case-out protocol. IBTS is only evaluated on the four models with a held-out split large enough to train the 3D model (Sup. F.2).

6.2 Calibration results

Does Ψ improve real model outputs without held-out labels? We evaluate whether the map improves calibration for inferred image patches (Table 1, block 1). Calibration improves on all ten models, and on both metrics. ECE falls by 1.3–$4.5\times$ and NLL by 1.2–$2.5\times$ on the ten models. NLL improves alongside ECE. Since a flattening towards the base rate can improve both, recovery of the posterior itself is tested where the posterior is known (Sup. D, Sup. E).

Do image-dependent f and D estimates improve compared to fixed values? By default, we read both scalars from the image by solving Eq. (8), then blend them through Eq. (9), for models where the Dice is pooled. Holding the functional form fixed and varying only f and D measures how much image dependence the real models carry. Replacing D_{eff} by the pooled constant $D^\star$ has relatively little impact on the organ models and decreases performance up to 0.04 of ECE on the lesion models (Table 1, block 1). Pinning both scalars yields the loss-optimal sample-agnostic pair Eq. (7) with a single fixed map per model, without image dependence. This single monotone map still cuts ECE by 1.2–$1.7\times$, but a per-image f improves on it on nine of ten models by each metric, most on Liver f and KiTS f.

How can both sources of miscalibration be corrected? At inference time, the miscalibration from imperfect generalization adds to the one induced by the loss. The miscalibration induced by the loss survives perfect training, and Eq. (6) removes it without held-out labels. The miscalibration from

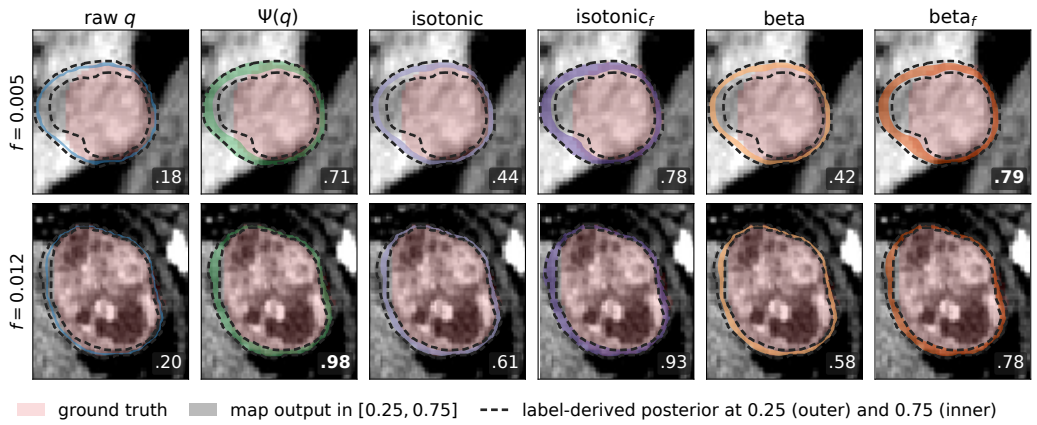

Figure 3: **Uncertainty margins on two KiTS lesions.** Our proposed maps better match the reference posterior contours. Each panel fills the voxels a map produces between the thresholds 0.25 and 0.75. The dashed contours mark the label-derived posterior at 0.25 and 0.75. The number in each image is the IoU of Table 2, computed on a window around the lesion. Rows: two lesions of different sizes, f the label fraction of the patch around them.

imperfect generalization arises because the model fits its training data rather than the population, and only held-out labels can reveal it (Sup. I). Block 3 of Table 1 shows that fitting an isotonic/beta map after Ψ beats fitting it directly in most cases. Isotonic regression benefits from the image dependence that Ψ provides, and beta calibration from both this image dependence and the functional form of Ψ.

Does conditioning the fitted maps on f improve compared to global maps? With abundant labels, a single map that jointly fits both sources of miscalibration may be the best configuration. We therefore condition the isotonic and beta maps on f. This improves on the unconditioned variants for both families, by NLL more consistently than by ECE (block 4, bold). IBTS (Ding et al., 2021), trained on the four splits large enough for it, shows mixed results, improving 2/4 NLL results and 2/4 ECE results.

More than one foreground channel. With C foreground channels, the soft Dice is averaged over channels and the softmax couples them, so the balance condition has one coupled solution, which reduces to Eq. (6) at $C = 1$ (Sup. J). On pancreas and liver ($C = 2$) and KiTS23 ($C = 3$) Ψ lowers NLL on all six columns by 14% to 58% (Sup. Table 6).

6.3 Uncertainty margins

A threshold on a probability map selects a region, and between two thresholds, it leaves a margin of voxels that are differently classified. We compare post hoc calibration methods using the IoU of the margin, ranging from outputs of 0.25 to 0.75, and the margin of the label-derived posterior at the same thresholds per patch across the six lesion models, where the margin represents a clinically relevant area. The margin from the raw output only yields an IoU of 0.10 to 0.18 (Table 2, Figure 3). The label-free map raises this to 0.34 to 0.47 and lies above the best image-free map fitted on held-out labels, on all six models (significantly on five). Conditioning on f raises the mean IoU on all six models for isotonic and on five for beta, significantly on four models for each family.

			one map per model		image-dependent		
model	n	raw	beta	isotonic	$\Psi(f_{\text{eff}}, D_{\text{eff}})$	isotonic$_f$	beta$_f$
KiTS l	141	$.14_{\pm.03}$	$.26_{\pm.03}$	$.30_{\pm.03}$	$\mathbf{.42}^{*}_{\pm.04}$	$.34_{\pm.03}$	$.35^{*}_{\pm.04}$
KiTS f	133	$.17_{\pm.02}$	$.28_{\pm.03}$	$.28_{\pm.03}$	$.34^{*}_{\pm.03}$	$\mathbf{.40}^{*}_{\pm.03}$	$.34^{*}_{\pm.03}$
Colon l	35	$.10_{\pm.03}$	$.32_{\pm.06}$	$.38_{\pm.07}$	$\mathbf{.43}_{\pm.08}$	$.39_{\pm.07}$	$.36_{\pm.07}$
Colon f	17	$.18_{\pm.08}$	$.26_{\pm.08}$	$.27_{\pm.09}$	$.45^{*}_{\pm.11}$	$\mathbf{.48}^{*}_{\pm.08}$	$.46^{*}_{\pm.10}$
Liver l	34	$.13_{\pm.03}$	$.38_{\pm.11}$	$.43_{\pm.11}$	$.47^{*}_{\pm.10}$	$\mathbf{.48}^{*}_{\pm.11}$	$.37_{\pm.09}$
Liver f	57	$.12_{\pm.03}$	$.22_{\pm.04}$	$.27_{\pm.04}$	$\mathbf{.47}^{*}_{\pm.06}$	$.40^{*}_{\pm.05}$	$.36^{*}_{\pm.05}$

Table 2: **Uncertainty margins.** IoU between the ring of voxels a map leaves between the thresholds 0.25 and 0.75 and the ring of the label-derived reference (Sup. G) on the lesion models, mean over patches with its case-level bootstrap standard deviation. * marks an image-dependent map significantly above its image-free counterpart, isotonic$_f$ against isotonic, beta$_f$ against beta and $\Psi(f_{\text{eff}}, D_{\text{eff}})$ against the better of the two. The conditioned maps condition on the per-patch f of Eq. (8). **Bold** highlights the best method.

7 Conclusion

Widely used compound segmentation losses distort predicted probabilities in ways that depend on the individual image, limiting what a shared calibration map can correct. We derive a simple image-specific map that improves calibration without retraining or held-out labels. When labels are plentiful, conditioning the fitted map on the image's own f improves the fit compared to the unconditioned case for most models. Using Ψ and the proposed image-conditioned post hoc calibration methods improves not only raw calibration metrics but also downstream tasks, such as detecting the label-derived 0.25–0.75 posterior probability margin.

7.1 Limitations

Our approach assumes that a network's outputs are at the stationary point of the loss for the image's own (f, D) (Remark 1). This requires the network to infer f and D from its input. Networks that can read a global statistic of their input track it where the posterior is known (Sup. D). On the real models, whose posterior is unknown, the evidence is indirect. The solved pair is close to the label-measured one (Sup. Table 3), and a per-image pair improves on a fixed one (Table 1). Ψ removes only the miscalibration induced by the loss. The miscalibration from imperfect generalization always requires held-out labels. At small f, the map compresses a wide band of outputs into a narrow band of posteriors, so errors in the output or in the estimated f are amplified (Sup. A).

Reproducibility statement

The models use nnU-Net with its planner configuration and are trained using the public `nnUNetv2` pipeline without modification. The five tasks are public Medical Segmentation Decathlon Task03 (liver, tumor label only), Task07 (pancreas, organ and tumor merged into one foreground class), Task09 (spleen), Task10 (colon), and KiTS23 (tumor label only) (Sup. F). Two schedules are used, $T = 1000$ and $T = 2000$ epochs. All reported numbers come from nnU-Net's fold 0. The evaluation region, the admission rules, and the leave-one-case-out protocol used are described in Sup. F. The code implementing the described methods will be released publicly upon acceptance.

Ethics statement

This work uses only publicly available, de-identified retrospective imaging datasets, under their respective licenses and intended research use.

AI USE STATEMENT

In this work, we used generative AI tools to generate the synthetic data sets with a known posterior (Sup. D and E), to formulate mathematical claims and provide ingredients for their proofs, to assist in writing the proofs (Sup. K), to co-develop the research methodology and the experiments, in particular those with a known posterior, to implement the methods, and to help develop parts of the theoretical model. Results were discussed and interpreted iteratively with AI support, and this joint discussion and interpretation led to a joint refinement of hypotheses. We have not used generative AI tools for translation or for the cleaning or reformatting of data sets, and qualitative or thematic data analysis does not apply to this work. Additionally, we used generative AI tools to design and revise figures. We have reviewed all AI-assisted work. Mathematical claims and interpretations of results were checked by the authors before being included in the paper. The code was tested for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Miriam Ayer, H Daniel Brunk, George M Ewing, William T Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, pp. 641–647, 1955.

Han Bao and Masashi Sugiyama. Calibrated surrogate maximization of linear-fractional utility in binary classification. In *International Conference on Artificial Intelligence and Statistics*, pp. 2337–2347. PMLR, 2020.

Theodore Barfoot, Luis C Garcia Peraza Herrera, Ben Glocker, and Tom Vercauteren. Average calibration error: A differentiable loss for improved reliability in image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 139–149. Springer, 2024.

Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4413–4421. IEEE, 2018.

Jeroen Bertels, David Robben, Dirk Vandermeulen, and Paul Suetens. Theoretical analysis and experimental validation of volume bias of soft dice optimized segmentation maps in the context of inherent uncertainty. *Medical Image Analysis*, 67:101833, 2021.

Noah Bice, Neil Kirby, Ruiqi Li, Dan Nguyen, Tyler Bahr, Christopher Kabat, Pamela Myers, Niko Papanikolaou, and Mohamad Fakhreddine. A sensitivity analysis of probability maps in deep-learning-based anatomical segmentation. *Journal of applied clinical medical physics*, 22(8):105–119, 2021.

Nontawat Charoenphakdee, Jayakorn Vongkulbhisal, Nuttapong Chairatanakul, and Masashi Sugiyama. On focal loss for class-posterior probability estimation: A theoretical perspective. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5198–5207. IEEE, 2021.

Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. Local temperature scaling for probability calibration. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6869–6879. IEEE, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, et al. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv preprint arXiv:2307.01984*, 2023.

Burhan Rashid Hussein, Cédric Meurée, Malo Gaubert, Arthur Masson, Anne Kerbrat, Benoît Combès, and Francesca Galassi. A study on loss functions and decision thresholds for the segmentation of multiple sclerosis lesions on spinal cord mri. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2023.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in neuroscience*, 14:282, 2020.

Viacheslav Komisarenko and Meelis Kull. Improving calibration by relating focal loss, temperature scaling, and properness. In *ECAI 2024: 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain–Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, pp. 1535–1542. SAGE Publications Pvt. Ltd 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2024.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pp. 623–631. PMLR, 2017.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in neural information processing systems*, 32, 2019.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Laurin Lux, Alexander H Berger, Moritz Knolle, Daniel Rückert, and Johannes C Paetzold. Beyond scalar losses: calibrating segmentation models via gradient vector field surgery. *arXiv preprint arXiv:2607.14338*, 2026.

Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878, 2020.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

Marcus Nordström, Han Bao, Fredrik Löfman, Henrik Hult, Atsuto Maki, and Masashi Sugiyama. Calibrated surrogate maximization of dice. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 269–278. Springer, 2020.

Marcus Nordström, Henrik Hult, Fredrik Löfman, and Jonas Söderberg. On image segmentation with noisy labels: Characterization and volume properties of the optimal solutions to accuracy and dice. *Advances in Neural Information Processing Systems*, 35:34321–34333, 2022.

Marcus Nordström, Henrik Hult, and Atsuto Maki. Marginal thresholding in noisy image segmentation. *arXiv preprint arXiv:2304.04116*, 2023a.

Marcus Nordström, Henrik Hult, Atsuto Maki, and Fredrik Löfman. Noisy image segmentation with soft-dice. *arXiv preprint arXiv:2304.00801*, 2023b.

Teodora Popordanoska, Jeroen Bertels, Dirk Vandermeulen, Frederik Maes, and Matthew B Blaschko. On the relationship between calibrated predictors and unbiased volume estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 678–688. Springer, 2021.

Mark D Reid and Robert C Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 11:2387–2422, 2010.

Axel-Jan Rousseau, Thijs Becker, Simon Appeltans, Matthew Blaschko, and Dirk Valkenborg. Post hoc calibration of medical segmentation models. *Discover Applied Sciences*, 7(3):180, 2025.

Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pp. 379–387. Springer, 2017.

Gihan Samarasinghe, Michael Jameson, Shalini Vinod, Matthew Field, Jason Dowling, Arcot Sowmya, and Lois Holloway. Deep learning for segmentation in radiation therapy planning: a review. *Journal of Medical Imaging and Radiation Oncology*, 65(5):578–595, 2021.

Michael Yeung, Leonardo Rundo, Yang Nan, Evis Sala, Carola-Bibiane Schönlieb, and Guang Yang. Calibrating the dice loss to handle neural network overconfidence for biomedical image segmentation. *Journal of Digital Imaging*, 36(2):739–752, 2023.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Supplementary Material

A Which losses admit a derived map?

Properness, of CE, MSE, and focal loss. If Ψ is the identity, the loss is proper, and outputs need no correction. By Eq. (3) this happens exactly when

$$\frac{g_0(q)}{g_1(q)} = \frac{q}{1-q} \qquad \text{for all } q \in (0,1). \tag{10}$$

This constrains only the ratio of the two pulls and does not restrict their magnitude, which is the distinction the criterion below rests on. Cross-entropy has $g_1 = 1/q$ and $g_0 = 1/(1-q)$, so the ratio is $q/(1-q)$, $\Psi = \mathrm{id}$, and its outputs are already posterior probabilities. Squared error, $\ell_1 = (1-q)^2$, has $g_1 = 2(1-q)$ and $g_0 = 2q$, the same ratio, and is also proper. Focal loss (Lin et al., 2017) has

$$g_1(q) = \frac{d_\gamma(q)}{q}, \qquad d_\gamma(q) = (1-q)^\gamma - \gamma(1-q)^{\gamma-1} q \log q, \tag{11}$$

whose ratio is not $q/(1-q)$. Substituting Eq. (11) into Eq. (3) reproduces the correction of Charoenphakdee et al. (2021).

The criterion. The derived map recovers η only where the risk of Proposition 1 has an interior minimizer. We call a loss *recoverable at* η when it does, and *fully recoverable* when it does at every $\eta \in (0,1)$. A loss failing the criterion may stay recoverable on part of the range, as MSE + Dice at $f = \frac{1}{2}$ does at half of all posteriors.

Assumption 1. *g_1 is continuous, positive and non-increasing; g_0 is continuous, positive and non-decreasing; and g_0/g_1 is strictly increasing.*

Every loss considered here satisfies Assumption 1, which makes the *pointwise* risk of Sec. 3.2 quasi-convex, so Ψ is strictly monotone and q^* is interior exactly on the open interval between the values of Eq. (3) at the two ends of the range.

Proposition 6 (The edges)**.** *Under Assumption 1, q^* is interior exactly when $\eta_{\mathrm{lo}} < \eta < \eta_{\mathrm{hi}}$, the values of Eq.* (3) *at the two ends of the range,*

$$\eta_{\mathrm{lo}} = \frac{g_0(0^+)}{g_1(0^+) + g_0(0^+)}, \qquad \eta_{\mathrm{hi}} = \frac{g_0(1^-)}{g_1(1^-) + g_0(1^-)}, \tag{12}$$

where 0^+ and 1^- denote the one-sided limits $q \downarrow 0$ and $q \uparrow 1$: the pulls need not be defined at the endpoints themselves, and CE's are infinite there.

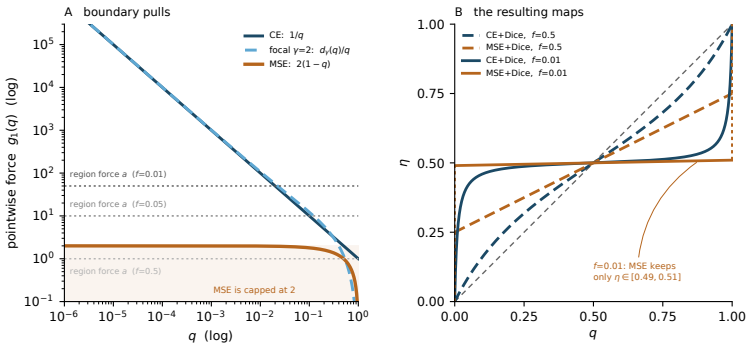

Figure 4: **Recoverable iff the pointwise pull diverges at the boundary.** **(A)** The upward pull against the constant region pull: CE and focal exceed any a, squared error caps at 2. **(B)** The resulting maps: MSE + Dice is linear on a window of width $2f/(2f+1)$ and dotted verticals mark posteriors it cannot represent. CE+Dice is fully recoverable in comparison.

The hypothesis of the theorem on the pointwise pulls is Assumption 1, and both results are proved in Sup. K.

Theorem 1 (Recoverability criterion)**.** *Let $L = L^{\mathrm{p}} + L^{\mathrm{r}}$ have finite region pulls $a, b > 0$ and continuous, positive pointwise pulls, g_1^{p} non-increasing and g_0^{p} non-decreasing with a strictly increasing ratio. Then L is fully recoverable if and only if*

$$\lim_{q\to 0^+} g_1^{\mathrm{p}}(q) = \infty \qquad \textit{and} \qquad \lim_{q\to 1^-} g_0^{\mathrm{p}}(q) = \infty. \tag{13}$$

By Eq. (12), a pointwise loss that pulls without bound as $q \to 0$ makes the denominator of η_{lo} diverge, so $\eta_{\mathrm{lo}} = 0$ and no posterior is lost. A bounded pull is eventually exceeded by the constant region pull, the optimum moves to the corner, and a whole range of posteriors produces the same output. Figure 4(A) contrasts the two cases. CE and focal both have $g_1^{\mathrm{p}} \sim 1/q$, and they pass the criterion combined with Dice, Tversky or Jaccard alike, since a region term changes only the two constants. MSE fails, its pull capping at 2, so (12) and (5) give

$$\Psi(q) = \frac{2q+b}{2+a+b}, \qquad \eta \in \left[\tfrac{b}{2+a+b},\ \tfrac{2+b}{2+a+b}\right], \tag{14}$$

a window of width $2/(2+a+b) = 2f/(2f+1)$ by (5). The foreground fraction alone sets the width, since $a+b = 1/f$, and therefore determines the size of the magnitude of the constants a and b. The window retains 50% of the posterior range at $f = 0.5$, 9% at $f = 0.05$ and 2% at $f = 0.01$ (Figure 4B), and such a model is a point predictor almost everywhere.

Beyond CE + Dice. Replacing the pointwise component with a different pointwise loss, or the region components with a different region component, preserves the same construction.

Focal + Dice. Focal loss has $g_1^{\mathrm{p}} = d_\gamma(q)/q$, where d_γ is its modulating factor Eq. (11). Putting this into Eq. (3) gives the same skeleton as Eq. (6), with $d_\gamma(q)$ in place of the constant numerator of CE's $1/q$ pull. Since $d_\gamma(0) = 1$ the boundary pull still diverges like $1/q$, so Theorem 1 applies.

Tversky and Jaccard. Repeating Eq. (4) for the Tversky loss (Salehi et al., 2017), with α_{T} weighting false negatives and β_{T} false positives, gives $a/b = \left[1/\mathrm{TI} - (1-\alpha_{\mathrm{T}})\right]/\beta_{\mathrm{T}}$, reducing to $(2-D)/D$ at $\alpha_{\mathrm{T}} = \beta_{\mathrm{T}} = \frac{1}{2}$ and to $1/J$ at $\alpha_{\mathrm{T}} = \beta_{\mathrm{T}} = 1$, the soft Jaccard behind Lovász-Softmax (Berman et al., 2018). Soft Dice and Jaccard both reach $a = b$ at a perfect fit, so their maps fix $\frac{1}{2}$ and a 0.5 threshold is correct for a converged model, whereas the asymmetric Tversky never does. At $\alpha_{\mathrm{T}} = 0.7$, $\beta_{\mathrm{T}} = 0.3$ the pulls stay in the ratio $7{:}3$ even at $\mathrm{TI} = 1$, so trading false negatives against false positives displaces the decision threshold *permanently*, not only while the model is imperfect.

The slope at the midpoint. The slope of the map at the midpoint,

$$\Psi'_{f,D}\left(\tfrac{1}{2}\right) = \frac{4f}{1+4f}, \tag{15}$$

is dependent solely on f and not on D. The map exists for every $f > 0$, but the slope falls with the object size. This means a wide band of outputs is compressed into a narrow band of posteriors, and inverting it can amplify any error in the output or in $\hat{f}$. In the limit, the map degenerates to a constant.

Proposition 7 (Degeneration)**.** *For fixed D, $\Psi_{f,D}(q) \to D/2$ pointwise on $(0,1)$ as $f \to 0$, and $\Psi'_{f,D}(\frac{1}{2}) \to 0$. The limit is a constant that is independent of q, and hence not invertible.*

B ESTIMATING (f, D)

This section derives the global pair of Eq. (7), shows that Eq. (8) has exactly one solution, and compares that solution with the pair the held-out labels measure.

B.1 THE GLOBAL PAIR $(f^\star, D^\star)$

A model that cannot estimate f and D from its input emits $q = \Phi(\eta)$ for one fixed map Φ, and Eq. (7) is the pair the loss-minimizing Φ corresponds to. Let $\rho_k(\eta)$ be the density of posterior values in training patch k. Since CE is a per-voxel mean, perturbing Φ at one value η affects $N\rho_k(\eta)$ voxels of patch k, each with the region pull $\frac{1}{N}[-\eta a_k + (1-\eta)b_k]$. The N cancels, and stationarity reads

$$\eta\Big(\tfrac{1}{\bar\Phi} + \bar a(\eta)\Big) = (1-\eta)\Big(\tfrac{1}{1-\bar\Phi} + \bar b(\eta)\Big), \qquad \bar a(\eta) = \frac{\sum_k \rho_k(\eta)\, a_k}{\sum_k \rho_k(\eta)}, \tag{16}$$

and likewise for $\bar b$. Each patch enters with the number of its voxels at posterior η, so the weights depend on η. For η near 1 these voxels are the object's interior, whose number is proportional to f_k. This weight cancels the $1/f_k$ in $a_k = (2 - D_k)/2f_k$, so as $\eta \to 1$, $\bar a \to (2 - \bar D)/2f^\star$, with $f^\star$ the mean patch fraction of Eq. (7) and $\bar D$ the mean patch Dice. Under a pooled Dice every a_k coincides and Eq. (7) is exact. Under a per-sample Dice, Eq. (7) keeps $f^\star$ and takes the pooled Dice $D^\star$ in place of $\bar D$, which direct minimization supports (below). Patches without foreground are excluded from K: with $S_k = 0$ the soft Dice is zero for any output, so they exert no region pull.

Minimizing the loss directly over synthetic patch distributions with median Dice 0.47 to 0.83 places the optimal D within 0.01 of $D^\star$, and under a pooled Dice reproduces $f^\star$ and $D^\star$ to four digits. Weighting every patch equally instead, $1/\mathbb{E}[1/f]$ and $\mathbb{E}[D/f]/\mathbb{E}[1/f]$, lets near-empty patches dominate. It places D 0.06–0.22 too low, and on spleen `3d_fullres` it returns $D = 0.64$ for a network whose median patch Dice is 0.98. Under a per-sample Dice no single $\Psi_{f,D}$ is exactly optimal, since the a_k differ between patches, but the best one changes ECE against Eq. (7) only in the fourth decimal on every model.

B.2 THE PER-IMAGE PAIR $(\hat f, \hat D)$

Since $\partial\Psi/\partial f \propto q - D/2$ changes sign, the right-hand side of Eq. (8) is not monotone in f, but the solution is still unique.

Proposition 8 (One root)**.** *For any outputs $q_i \in (0, 1)$, the system* (8) *has exactly one solution with $f > 0$ and $D > 0$, and it has $\bar q/2 \le f \le (\bar q + 1)/2$ and $D \le 1$.*

Read as curves $D(f)$ in the (f, D) plane, the first equation is strictly increasing wherever it is positive and the second strictly decreasing, so they cross once (§K). There is no root to select, and the solve does not depend on its starting point.

The solved pair against held-out labels. A unique solution need not be the right one: Eq. (8) puts the map's own output where the labels should be. Its two equations are identities that held-out labels can check, $\sum_i \Psi(q_i) = S_Y$ and $\sum_i q_i\Psi(q_i) = I_Y$, that is, $\hat f = f_Y$ and $\hat D = D_Y$ for the pair the labels measure. Table 3 compares them per whole validation patch on the 746 patches of Table 1 that contain label foreground. The median $\hat f/f_Y$ is within 2% of one on six models and between 0.94 and 0.96 on three; on colon `3d_lowres` it is 0.85, where the model's own soft volume is 0.80 of the label volume. $\log \hat f$ tracks $\log f_Y$ at $r = 0.90$–1.00. The median $\hat D$ is within 0.06 of D_Y on all models but colon `3d_lowres` (0.09), closer on every model than the soft Dice of the output with itself, which overestimates D_Y by 0.02 to 0.21, but it correlates with D_Y only at 0.38–0.98. D is identified weakly because its coefficient $\partial\Psi/\partial D = q(1-q)/2[f + q(1-q)]$ vanishes as $q \to 0$ or $q \to 1$: $\hat D$ is least certain where the map depends on it least.

	Panc. l	Panc. f	Spleen l	Spleen f	KiTS l	KiTS f	Colon l	Colon f	Liver l	Liver f
patches	240	92	27	23	120	101	35	14	44	50
median $\hat{f}/f_Y$	0.98	1.00	0.99	0.99	1.00	1.00	0.85	0.94	0.96	0.94
$r(\log \hat{f}, \log f_Y)$	0.97	0.97	1.00	0.99	0.99	0.90	0.94	0.97	0.95	0.99
median $\hat{D} - D_Y$	+.043	+.042	-.002	+.003	-.016	+.014	+.091	+.058	+.054	+.028
median $D_{qq} - D_Y$	+.115	+.112	+.021	+.018	+.088	+.065	+.211	+.131	+.155	+.124
$r(\hat{D}, D_Y)$	0.38	0.63	0.97	0.93	0.70	0.79	0.70	0.98	0.70	0.54

Table 3: The pair solved from the output by (8) against the pair the held-out labels measure, $f_Y = (P + S_Y)/2N$ and $D_Y = 2I_Y/(P + S_Y)$, per whole validation patch, over the patches Table 1 scores that contain label foreground. The estimator never sees S_Y or I_Y. $D_{qq} = \sum_i q_i^2 / \sum_i q_i$ is the soft Dice of the output against itself, the label-free guess at D that uses no map.

C THE POOLED-DICE EFFECTIVE PAIR

Under `batch_dice=True`, the effective pair of Eq. (9) replaces the image's own pair (Section 4). This section derives the effective pair, verifies it on networks trained at different batch sizes, and compares the two-step estimator of the tables with the coupled one.

Derivation. Under a pooled Dice a batch of B patches is one concatenated image with $P + S = 2N \sum_k f_k$, so the region pull on a voxel scales as $1/\sum_k f_k$. CE is a mean over all BN voxels of the batch and carries a factor of $1/(BN)$ per voxel. The stationarity of CE + Dice at a voxel with posterior η then reads

$$(q - \eta) \;=\; \frac{BN \cdot 2}{2N \sum_k f_k} \left(\eta - \tfrac{D}{2}\right) q(1-q) \;=\; \frac{1}{f_{\text{avg}}} \left(\eta - \tfrac{D}{2}\right) q(1-q), \tag{17}$$

The displacement is that of a single image of fraction f_{avg}. Under `batch_dice=False` the Dice is a per-sample mean whose own $1/B$ cancels that of cross-entropy, leaving f_k per image. The split of the pooled pull between the two branches is set by D_{batch} at the same weight $1/(P{+}S) \propto 1/f_{\text{avg}}$. Averaging both over the partners the sampler pools with gives Eq. (9). Since $1/f_{\text{avg}}$ is convex, f_{eff} lies between the harmonic mean of the patch fractions, which it approaches at small B, and their arithmetic mean, which it approaches as B grows.

Verification. We train one network per architecture and batch size on a mixture spanning a $29\times$ range of f, with the range varying with object count. Varying by object count makes the local evidence at a boundary independent of f. Then we recover the effective fraction each network converged to as the (f, D) at which Ψ reproduces the true posterior (Figure 5). Eq. (9) tracks it with an RMSE of about 0.005 at every batch size, against 0.006–0.031 for the harmonic and 0.008–0.015 for the arithmetic mean. The fitted weight of the image's own f scales as $1/B$, with $wB \approx 1.2$ against the 1.1 that Eq. (9) predicts. Both results hold for a 31k-parameter network that reads f only through its normalization statistics and for a 1.4M-parameter U-Net.

The coupled solve. The tables solve Eq. (8) with $\Psi_{f,D}$ and blend the result once by Eq. (9). The coupled estimator instead lets $\Psi_{f_{\text{eff}}, D_{\text{eff}}}$ stand in for the image's labels in Eq. (8), with $(f_{\text{eff}}, D_{\text{eff}})$ computed from those moments and the training partners by Eq. (9). Starting with the two-step solution, a root-finder converges on all 313 evaluation patches across the five pooled models. The coupled solution moves f_{eff} by a median of under 2% and D_{eff} down by 0.01–0.04.

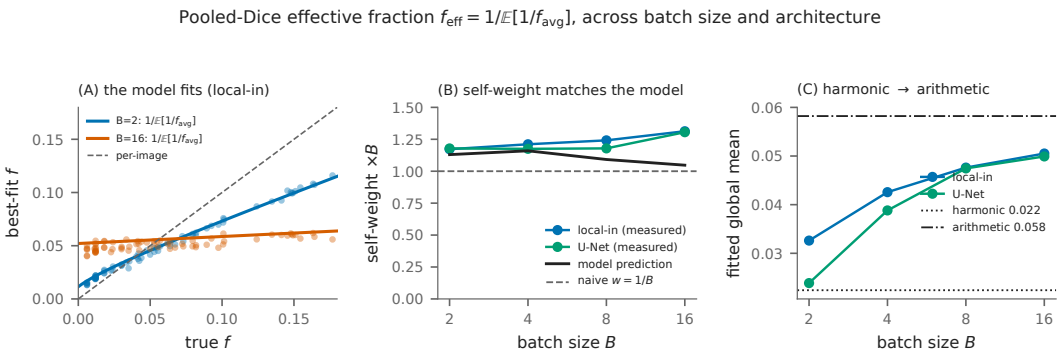

Figure 5: **The pooled-Dice effective fraction. (A)** f_{eff} of Eq. (9) (curves) against the fraction each trained network converged to (points), at $B = 2$ and $B = 16$. **(B)** The measured weight of the image's own f times B (points) against the prediction of Eq. (9) (line), for both networks. **(C)** The fitted global fraction moves from the harmonic mean at $B = 2$ toward the arithmetic mean as B grows.

D SYNTHETIC EXPERIMENTS WITH A KNOWN POSTERIOR

Across the ten trained models, f moves together with anatomy, resolution and voxel size, so they cannot show that f causes the displacement. Three manipulations with a known posterior each vary one quantity: f across networks (§D.1), f within one network (§D.2), and D within one network (§D.3).

The construction. The posterior is known because the uncertainty is placed in the labels rather than the input. The input is a deterministic, noiseless transform of the distance $d(v)$ to an object's centre, the posterior is $\eta(v) = \sigma(\kappa(R - d(v)))$, and each label is an independent draw $y(v) \sim \text{Bernoulli}(\eta(v))$. A network trained with CE alone would therefore converge to η, and any systematic departure of a CE + Dice model from η is due to the region term. The loss is nnU-Net's `DC_and_CE_loss` with `do_bg=False`. At fixed boundary softness κ a smaller object has more boundary per unit volume, so the attained Dice falls from 0.902 at $f = 0.5$ to 0.667 at $f = 0.005$. Every prediction is stated against the (f, D) measured per image.

D.1 ONE NETWORK PER f

We train one network per foreground fraction, eight values from 0.005 to 0.5, with four seeds each and eight held-out volumes per run (Figure 6). *(i) Recovery of f and D.* Averaged over the 32 volumes of each f, the label-free estimates of Eq. (8) track the truth at $r = 1.00000$ for f and $r = 0.99989$ for D, with the mean $\hat{D}$ never off by more than 0.0033. The one sizeable departure is $\hat{f}$, 11.3% low at $f = 0.005$ and 7.2% at $f = 0.01$ against at most 2.8% for every $f \geq 0.02$, the $f \to 0$ regime of Prop. 7. *(ii) Recovery of the posterior.* $|\Psi_{f,D}(q) - \eta|$ is 2.4 to $2.7\times$ smaller than $|q - \eta|$ for every $f \leq 0.05$, and the gain decays with f, to $1.84\times$ at $f = 0.1$, $1.27\times$ at 0.2 and $1.00\times$ at $f = 0.5$. *(iii) The decision threshold.* The output at which the posterior crosses $\frac{1}{2}$ correlates with the predicted threshold q^* of Sec. 3.3 at $r = 0.97$ but lies below it by 0.097 on average, at every f, with the largest gap in the middle of the range.

D.2 ONE NETWORK ACROSS MANY f

One network is trained with a per-sample Dice on a mixture whose foreground fraction varies over $27\times$, from 0.0060 to 0.1642. The fraction is varied by object *count*, so the local evidence at any boundary is the same whatever the image's f, and only a network with access to a global property of the image can condition on it. We compare three networks: a U-Net (1.40M parameters) with a global receptive field; a two-layer 3^3 convolution stack with instance normalization (31k), which sees a 5^3 neighbourhood but divides every voxel by a statistic of the whole volume; and a wider copy of the stack without normalization (67k) as the local control.

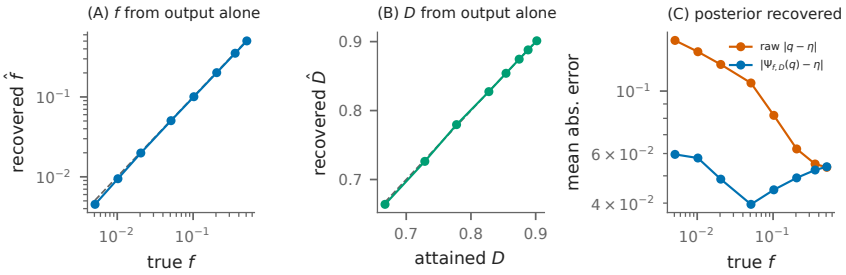

Figure 6: **One network per f. (A, B)** The label-free $(\hat{f}, \hat{D})$ against the truth. **(C)** Posterior error of the raw and the corrected output.

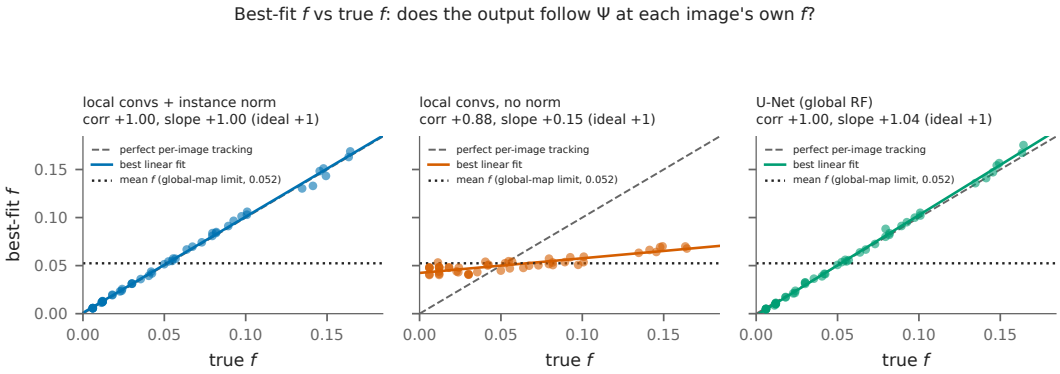

Figure 7: **One network across many f.** Best-fit f against each held-out image's true f, for one network per architecture trained on the $27\times$ mixture. Dotted line: the mean f a single global map converges to.

For each held-out image we fit the (f, D) at which Ψ best reproduces the true posterior and plot it against the image's true f (Figure 7). The normalized network and the U-Net follow the diagonal ($r = 0.999$, slopes 1.00 and 1.04). The best-fit f of the network without normalization rises with a slope of only 0.15 around a mean of 0.051, the training mean being 0.052. Against one global pair, a per-image (f, D) cuts the posterior error from 0.0544 to 0.0059 for the normalized network ($9.3\times$) and from 0.0604 to 0.0168 for the U-Net ($3.6\times$), but raises it from 0.0170 to 0.0580 for the control ($3.4\times$ worse). The estimator reads each image's f correctly on all three networks, the control included ($r \geq 0.9998$). The per-image link, therefore, requires access to a global statistic, which instance normalization supplies, and a U-Net has through its bottleneck.

D.3 One network across many D

To isolate D, the object count is adjusted per image to hold the foreground fraction at 0.047 ± 0.003, while the boundary softness varies over $\kappa \in [0.61, 5.85]$, so the attained Dice ranges from 0.51 to 0.93. The best-fit D follows the diagonal for all three networks (Figure 8; $r = 0.995$ to 0.997, slopes 0.96 to 1.01), the label-free estimate recovers it ($r \geq 0.999$), and a per-image D instead of one global D cuts the posterior error by $1.4\times$ for the U-Net, $4.2\times$ for the normalized and $5.3\times$ for the local network, within 1 to 18% of an oracle per-image D.

Scope. The three manipulations show that f governs the displacement, that a single network realizes it per image whenever it can read a global statistic, and that D is recoverable per image. They do not establish external validity, since the objects are spheres and the only irreducible uncertainty is label noise; the trained models of Sec. 6 cover real anatomy.

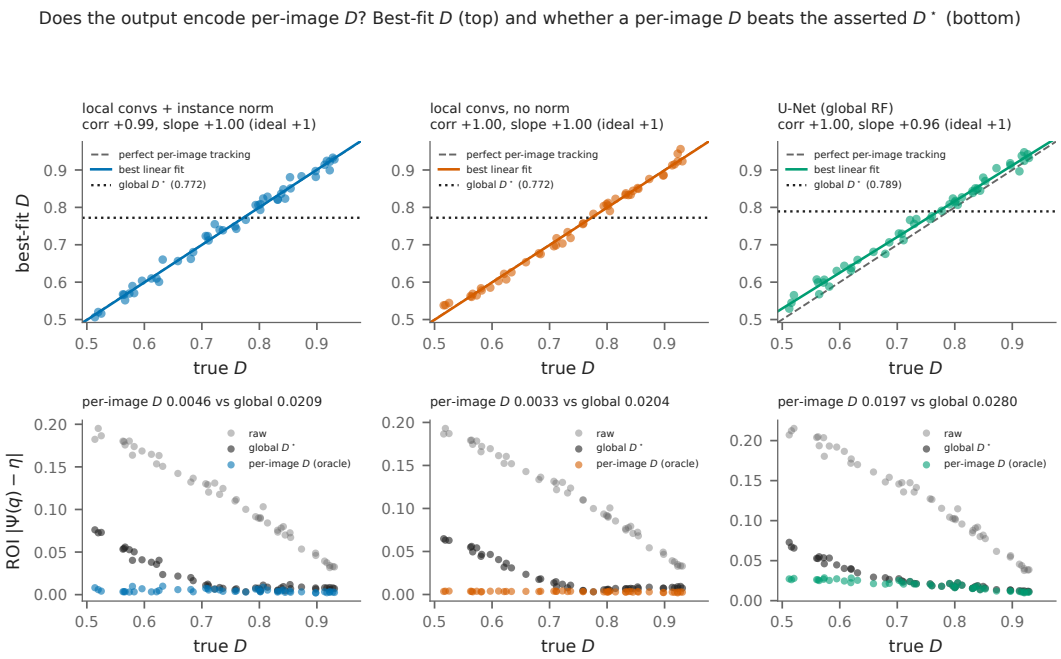

Figure 8: **One network across many D.** One network per column, trained with the foreground fraction held at 0.047 ± 0.003 while the boundary softness varies. **Top:** best-fit D against the image's true D; dotted line: the single global $D^{\star}$. **Bottom:** posterior error against true D for the raw output and the output corrected with $D^{\star}$ and with a per-image D.

E When the mask's volume fluctuates

Eq. (5) evaluates the region pulls $a(Y)$ and $b(Y)$ at the mean sums $\bar{S} = \mathbb{E}[S_Y \mid x]$ and $\bar{I} = \mathbb{E}[I_Y \mid x]$, whereas Proposition 2 makes them conditional means over the masks an image can carry. This section derives the difference to first order and measures the error of the map under a coherent fluctuation, against a known posterior.

The first-order departure. The plug-in is exact when S_Y and I_Y do not vary given x, and when the Dice is computed against η in place of Y. Otherwise, expanding $a(Y)$ and $b(Y)$ to first order about $(\bar{S}, \bar{I})$ and conditioning on Y_j, with $\mathbb{E}[S_Y - \bar{S} \mid Y_j = 1] = \mathrm{Cov}(S_Y, Y_j)/\eta_j$, $\mathbb{E}[S_Y - \bar{S} \mid Y_j = 0] = -\mathrm{Cov}(S_Y, Y_j)/(1-\eta_j)$ and likewise for I_Y, gives

$$\begin{aligned} \frac{a_j}{a} - 1 &= -\frac{(1-D)\,\mathrm{Cov}(S_Y, Y_j) + \mathrm{Cov}(I_Y, Y_j)}{\eta_j\,(P+\bar{S})\,(1-D/2)}, \\ \frac{b_j}{b} - 1 &= \frac{D\,\mathrm{Cov}(S_Y, Y_j) - \mathrm{Cov}(I_Y, Y_j)}{(1-\eta_j)\,(P+\bar{S})\,D/2}. \end{aligned} \tag{18}$$

Under conditionally independent labels $\mathrm{Cov}(S_Y, Y_j) = \eta_j(1-\eta_j)$ and $\mathrm{Cov}(I_Y, Y_j) = q_j\,\eta_j(1-\eta_j)$, so both departures are $O(1/(P+\bar{S}))$, and so are the second-order terms, since $\mathrm{Var}(S_Y \mid x) \le \bar{S}$. Under a coherent shift of the contour both covariances are of order $\rho\,\bar{S}$ for a voxel on the contour, and the departure is first order in the relative spread $\rho = \mathrm{sd}(S_Y)/\bar{S}$.

Construction. One sphere of radius R in a 48^3 volume, with the input a noiseless transform of the signed distance $R - d$ to its surface, so that each volume's boundary is visible to the network. The label thresholds a latent field, $Y_i = \mathbf{1}\{R - d_i + \delta + \epsilon_i > 0\}$, with $\delta \sim \mathcal{N}(0, \tau^2)$ drawn once per mask and $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ per voxel. The posterior is $\eta_i = \Phi((R - d_i)/\sigma)$ with $\sigma^2 = \tau^2 + \sigma_\epsilon^2$ for any split, so moving the coherent share $\lambda = \tau^2/\sigma^2$ from 0 to 1 at fixed σ leaves η, (f, D) and the target map unchanged and only changes how the mask's volume fluctuates, from independent voxels to a rigid shift of the contour. The relative spread ρ reaches 0.48 on the smallest object at a contour width of $\sigma = 1$ voxel and 0.77 at $\sigma = 1.6$.

Arms. The stationary point is reached in two ways: by a network trained with CE and per-sample soft Dice on a fresh mask at every step, and, for one held-out volume per setting, by one free logit per voxel, optimized directly and therefore at the stationary point of the conditional risk by construction. Against η on the evaluation region we score the raw output, $\Psi_{f,D}$ at the label-free pair of Eq. (8) and at the plug-in pair of Eq. (5), and, as references, the best pull pair (a, b) fitted to η and the exact link of Proposition 2, its pulls estimated by Monte Carlo over 4,096 masks.

Results. At the stationary point (Figure 9A), the plug-in map is exact for independent labels, reducing the error from 0.04–0.18 to 0.001–0.002. Its error grows with ρ, to 0.020 at $\rho = 0.48$ and 0.045 at $\rho = 0.77$, while the exact link stays at 0.001–0.003, so the whole departure is the plug-in's. The plug-in still removes 85 to 89% of the raw error at $\rho \approx 0.5$ and 77% at $\rho = 0.77$. The upward pull falls and the downward pull rises by nearly equal fractions, 0.12 each at $\rho = 0.48$, as the first-order expansion Eq. (18) predicts, so a coherent fluctuation changes the split of the pull but not its total: it acts like an effective Dice above the attained one and leaves f intact, and Ψ depends on D only weakly.

On the trained network (Figure 9B), the error after the plug-in map is 0.002–0.011 at $\rho \approx 0$, as for a network trained against the marginal η as a soft target, for which the plug-in is exact. It rises with ρ somewhat faster than at the stationary point, most on the largest object (0.0135 against 0.0018 at $f = 0.2$ and $\rho = 0.17$), and the exact link does not remove it, so the rise is the network's remaining distance from the stationary point. At a quarter of the schedule the network's error is 1.3 to 3.5 times as large, so this distance shrinks with training.

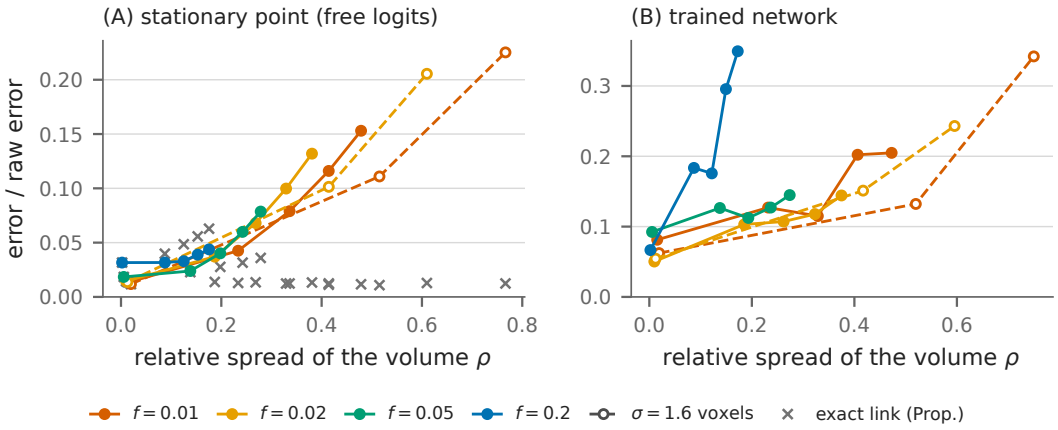

Figure 9: **The map under a mask whose volume fluctuates.** Error of $\Psi_{f,D}$ at the label-free pair as a fraction of the raw output's error, against the relative spread ρ of the volume, one colour per object size. **(A)** The free-logit stationary point, with the exact link of Proposition 2 as a reference. **(B)** The trained network, mean over two seeds and eight held-out volumes. Open markers: contour width $\sigma = 1.6$ instead of 1 voxel.

F PROTOCOL AND BASELINES

Tasks and training. The map of Sec. 3.3 is derived for a single foreground channel, so each dataset is reduced to one binary task before training. Spleen (MSD Task09) and colon (MSD Task10) are binary as published. Pancreas (MSD Task07) includes an organ and a tumor label, which are merged into a single foreground class. KiTS23 ships kidney, tumor, and cyst, and liver (MSD Task03) ships liver and tumor. For these two datasets, only the tumor label is retained, resulting in small objects and a wide range of f. §J treats the native multi-class versions. Every model is nnU-Net v2 on fold 0 with the planner's configuration and default recipe, without early stopping and with inference from the final checkpoint. Of the two budgets run, 1000 and 2000 epochs, the one with the better final EMA pseudo-Dice is reported: 2000 epochs for KiTS23 and liver, 1000 for spleen, pancreas, and colon.

Deep supervision. nnU-Net's default trainer adds the loss at five decoder resolutions, $L = \sum_s w_s L_s(p^{(s)}, y^{(s)})$, on separate output tensors. The full-resolution output $p_j^{(0)}$ enters L_0 alone, so $\partial L/\partial p_j^{(0)} = w_0\, \partial L_0/\partial p_j^{(0)}$, and the positive constant w_0 divides out of Eq. (2) as the $1/N$ does. The map is therefore set by the output head's own (f, D), which is the head nnU-Net returns at inference.

Patches and scoring. Since the loss operates on patches, we evaluate the map at the patch scale. Each validation volume is tiled into non-overlapping patches, (f, D) is solved from each whole patch by Eq. (8), and metrics are computed on the central half of each tile, within the evaluation region $\mathrm{dilate}(\{q > \frac{1}{2}\}, r{=}3)$ (§F.1). A patch is scored if the model predicts at least 20 voxels above $\frac{1}{2}$ and the ROI contains at least 20. Both rules depend on the output, not on the labels, so every row of Table 1 is scored on the same voxels.

Table 1. $\Psi(f^\star, D^\star)$ is one fixed map per model at the pair of Eq. (7). $\Psi(f_{\text{eff}}, \cdot)$ uses the per image solved f, blended by Eq. (9) where the Dice is pooled, and $D^\star$ or D_{eff}. A subscript f marks the image-conditioned fit of Sec. 5, with the bandwidth $h = 1$ fixed for every model rather than a tuned bandwidth. A star marks $p < .01$ for the comparison the bold asserts, against the same partner on the same cases: an exact sign-flip permutation on the per-case differences for NLL, and a paired bootstrap over cases for ECE, which is a pooled statistic rather than a per-case mean. The first two blocks carry no stars, since their mark is the winner of a multi-way comparison. Of the four rows below the methods, the hard Dice is that of $\{q > \frac{1}{2}\}$ within each case over the evaluation regions of its scored patches, averaged over the cases with label foreground there, $\mathrm{sd}\log f$ is taken over the patches nnU-Net's sampler draws (Table 4), and the last two rows count the cases on which the label-free map lowers ECE and NLL against the raw output. There, ECE is the debiased squared calibration error of each case at the table's 15 bins (Kumar et al., 2019), $\sum_b (n_b/n)\,[(a_b - c_b)^2 - a_b(1-a_b)/(n_b-1)]$ with n_b voxels, mean output c_b and observed frequency a_b in bin b, since the plain per-case ECE is biased upward at the few hundred voxels of a lesion case.

measure	Spleen l	Spleen f	Pancreas l	Pancreas f	Colon f	Colon l	KiTS f	KiTS l	Liver l	Liver f
over the patches nnU-Net's sampler draws — what the loss met										
n	264	228	363	268	439	538	434	564	519	430
sd log f	0.442	0.452	0.479	0.509	1.027	1.099	1.556	1.759	2.023	2.158
over the evaluation patches — what the tables are scored on										
n	20	20	190	61	10	24	85	93	22	42
p_{90}/p_{10}	4.5×	3.5×	3.1×	3.4×	3.6×	15.7×	34.4×	111.3×	35.4×	20.5×

Table 4: How much the per-patch foreground fraction varies within each model. Upper block: the patches nnU-Net's sampler draws in training, the pool the loss met; Table 1 quotes its sd log f, and columns are ordered by it. Lower block: the evaluation patches, restricted for this measurement to those with at least 200 true foreground voxels, and the ratio of the 90th to the 10th percentile of their f.

Baselines. The fitted baselines of Table 1 are fitted leave-one-case-out on the evaluation labels. The comparison is not information-matched: the derived map uses no labels. With enough labels, a fitted map should win, §H measures how many labels are needed.

F.1 THE EVALUATION REGION

We use $\mathrm{dilate}(\{q > \frac{1}{2}\}, r{=}3)$, selected from the model's output as the evaluation region for the calibration metrics.

Why evaluation is not done on all voxels. The foreground occupies a small fraction of the scans, and calibration metrics are averaged over voxels. Scored over the whole volume, the abundant background voxels decide the number. A model can be arbitrarily wrong wherever the decision is actually difficult and still post an excellent ECE, because the confidently empty voxels that dominate the average are dominating the results. Restricting to a region around the model's predicted foreground removes that irrelevant background mass. The region is a function of the output alone, so it conditions on the model but not on the labels.

F.2 IMAGE-BASED TEMPERATURE SCALING AS A BASELINE

Ding et al. (2021) predict a temperature per image (IBTS) with a network that reads the logit map and the image and is trained by NLL on a held-out split.

Protocol. The regressor takes the foreground logit map and the preprocessed image of a validation patch, average-pools by two, passes four stride-two convolution blocks of 16, 32, 64 and 64 channels with instance normalization, global-average-pools, and maps through a two-layer head to $\log T$, zero-initialized so that training starts at the raw output. It is trained using the NLL of $\sigma(z/T)$ over the evaluation region of each patch, with random flips and Adam at 10^{-3} under a cosine schedule, and the checkpoint is chosen from a fifth of the training cases held back. Like all the other fitted rows of Table 1, it is trained in leave-one-case-out style on the fold-0 validation cases, one network per held-out case. It is run only on the four models with enough cases in the validation split (46 to 85 cases). We train IBTS on the evaluation region rather than on every voxel of the patch. Trained on every voxel (as in Ding et al.), it is worse than a global temperature on all four models.

Details on the results. Its predicted temperatures do not track the per-patch optimum defined by the patch's own labels: over held-out patches, the correlation of the two $\log T$ is -0.03, -0.27, $+0.10$ and $+0.19$ on pancreas `3d_lowres`, pancreas `3d_fullres`, KiTS `3d_lowres` and KiTS `3d_fullres`, and between the 5th and 95th percentile of patches its output spans a factor of 1.3 to 2.6 where the optimum spans 4 to 80. At 46 to 85 cases it recovers a global temperature with a per-patch jitter, as Ding et al. (2021) report for their own 3D experiment on 40 brain MRI scans.

G UNCERTAINTY MARGINS

Between the thresholds 0.25 and 0.75, a margin of undecided voxels is created. Within a patch, every map of Table 2 is monotone in q, so its margin is an interval of the voxels ranked by q, and so is the margin of the posterior.

The score. To estimate the posterior's margin from the labels, the voxels of each patch are put into equal-count bins of at least 200 by q, each with a Wilson 95% interval for its label frequency, and the number of voxels above each level is bracketed between the highest bin below the level and the lowest above it. A map's margin is compared with this reference by their intersection over union, with each end of the reference taken at the bracket end nearest the map's, so a margin that lies within the bracket scores one. The score is computed on $\mathrm{dilate}(\{q > 0.01\}, r{=}3)$, where the count above every threshold used is exact, and averaged over patches. Table 2 reports its mean with a case-level bootstrap standard deviation for the raw output, the label-free map and the fitted maps of Table 1, all leave-one-case-out. A star marks an image-dependent map whose per-patch difference to its image-free counterpart has a case-level bootstrap 95% interval above zero ($2{,}000$ draws).

Results. On the lesion models, the best single fitted map is worse than the best image-dependent map by 5 to 21 points. Among the image-dependent maps, the label-free map and the f-conditioned isotonic provide the best results.

H HOW MANY LABELS A FITTED MAP NEEDS

We test how many labeled cases a fitted map needs to beat Ψ, which uses none, at NLL and ECE. For Figure 10, each fitted family is fitted on k labeled cases while the evaluation cases stay fixed. With every case weighted equally, the best of the four fitted families first beats Ψ by ECE after 1 to 12 labeled cases on seven of the ten models, only with the whole fitting pool (46 and 57 cases) on the two KiTS models, and at no budget on liver `3d_fullres`; by NLL it does so after 1 to 12 cases on nine models and after 32 on KiTS `3d_fullres`.

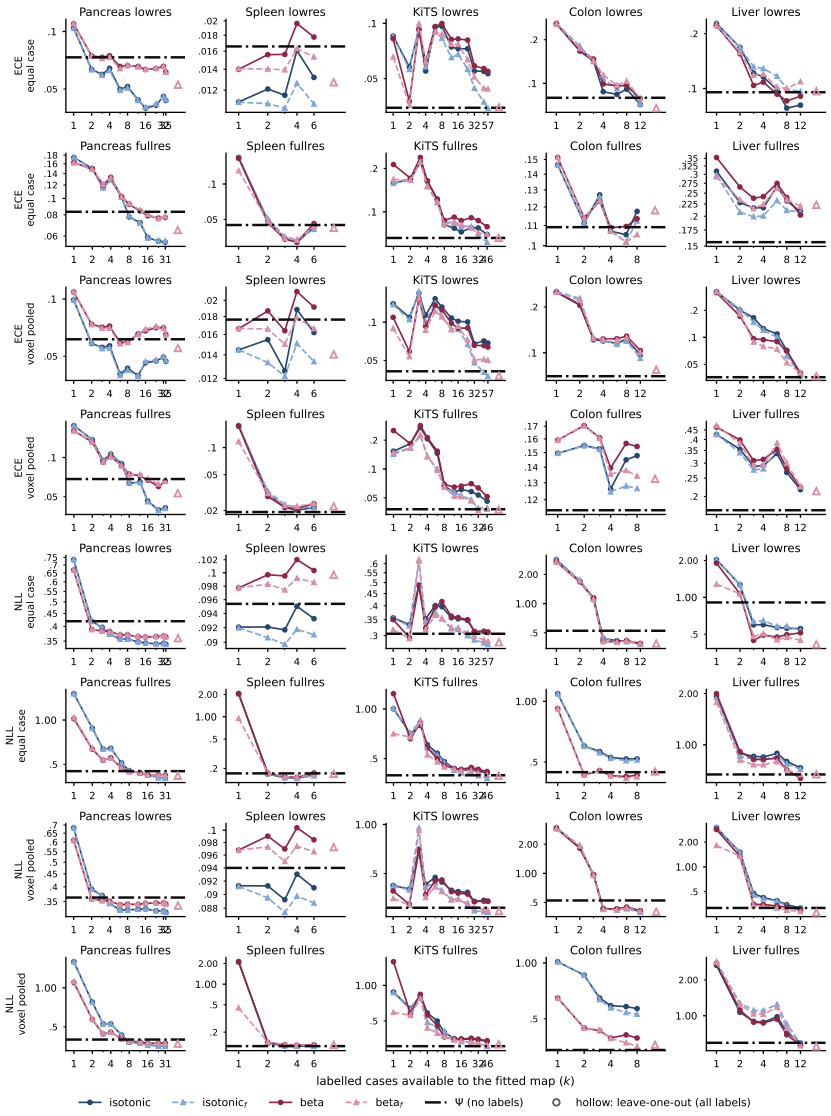

Figure 10: **ECE and NLL against the label budget.** Rows: metric, weighting (every case equal, or voxels pooled) and configuration; columns: dataset. A third of the cases is held out for evaluation at every k. The curves are means over five draws. Ψ (dash-dot) uses no labels. Isotonic and beta are depicted as solid lines. The f-conditioned variants are lighter and dashed. Hollow markers are the same families fitted leave-one-case-out on all labels evaluated on the same cases. Both axes are logarithmic.

I THE RESIDUAL THE MAP LEAVES ON TRAINING AND HELD-OUT PATCHES

We continue the finished training of two per-sample-Dice lesion models, KiTS l and Liver l, for 500 steps at the final learning rate, keep every batch the optimizer steps on, and score the network on exactly those batches afterwards. Each voxel is mapped with the $\Psi_{f,D}$ of its own sample, (f, D) taken from the sample's labels, and the binned label frequency is compared with the binned prediction on whole patches (Table 5). On these training samples, the per-sample map leaves 0.36 and 0.38 logits of the raw 2.63 and 3.63, where one pooled (f, D) leaves 2.09 and 2.20. The gap grows with f, and on the largest third of KiTS the pooled map is no better than the raw output (2.13 against 2.21). On the smallest third, lesions of a median 638 and 129 voxels, the pooled map is slightly ahead (0.76 against 0.89, 1.21 against 1.76). Batches drawn afresh from the training cases agree

with the trained-on ones to 0.01. On held-out samples the per-sample map, still with the labels' (f, D), leaves 0.99 and 2.46: this part exists only out of sample, grows with f, and is what a fitted rescaling has to remove (Sec. 6.2).

		$\Psi_{f,D}$ per sample				one pooled (f, D)				raw output			
		all	s	m	l	all	s	m	l	all	s	m	l
KiTS l	trained on	0.36	0.89	0.50	0.33	2.09	0.76	1.18	2.13	2.63	5.70	4.85	2.21
	held out	0.99	0.68	0.99	2.00	1.38	0.93	0.78	1.53	5.07	6.18	5.68	5.10
Liver l	trained on	0.38	1.76	0.85	0.44	2.20	1.21	1.36	2.20	3.63	7.08	6.49	3.17
	held out	2.46	1.09	2.41	4.60	2.78	2.02	1.77	3.47	8.80	8.80	8.87	9.55

Table 5: RMSE of $\operatorname{logit} \mathbb{E}[Y \mid q]$ against each prediction (logits, bins weighted by inverse variance), over all samples and by thirds of the label foreground fraction (s, m, l). Median f per third: 0.0003, 0.0015, 0.012 (KiTS) and 0.000062, 0.00069, 0.0071 (Liver).

J More than one foreground channel: the link at $C > 1$

The loss. nnU-Net builds `DC_and_CE_loss` with `do_bg=False` and unit weights, and its soft Dice ends with a mean over the C foreground channels. The region term of channel c is therefore scaled by $1/C$, with $a_c = (2 - D_c)/2Cf_c$ and $b_c = D_c/2Cf_c$, and the background channel carries no region pull.

The coupled map. The softmax couples the channels, so the map is not Ψ_{Cf_c,D_c} applied per channel: the balance condition Eq. (2) becomes one stationarity over the simplex. With $\lambda_0 = 0$ and $\lambda_c = (D_c/2 - \eta_c)/(Cf_c)$, the per-voxel risk $-\sum_k \eta_k \log p_k + \sum_k \lambda_k p_k$ is stationary at $p_k = \eta_k/(\lambda_k - \mu)$, with μ the multiplier of $\sum_k p_k = 1$. Solving for η gives the map $\boldsymbol{\Psi}\colon \mathbf{p} \mapsto \boldsymbol{\eta}$,

$$\Psi_0(\mathbf{p}) = -\mu\, p_0, \qquad \Psi_c(\mathbf{p}) = \frac{p_c \left[\frac{1}{2} D_c - \mu\, C f_c\right]}{C f_c + p_c}, \qquad \mu = \frac{B - 1}{A}, \tag{19}$$

with $A = p_0 + \sum_c C f_c\, p_c/(C f_c + p_c)$ and $B = \sum_c \frac{1}{2} D_c\, p_c/(C f_c + p_c)$. Since (f_c, D_c) are constants of the image, $\sum_k \Psi_k = 1$ is linear in μ, so μ has a closed form. At $C = 1$, Eq. (19) is Eq. (6).

Feasibility. An interior stationary point needs every $\Psi_k > 0$, that is $\mu < 0$, or $B < 1$. Since $p_c/(C f_c + p_c) < 1$, $B < \sum_c D_c/2 \le C/2$, so the condition always holds for $C \le 2$. At $C = 3$ it fails on 0.0004% and 0.003% of the scored voxels of the two KiTS models, where the output is left uncorrected.

The estimator. The divisor is the nominal C: the channel mean is unconditional, so a class absent from the patch exerts no pull but still counts. As in §4, f_c is the patch's own $(P_c + S_c)/2N$ under a per-sample Dice and the batch-effective f_c of Eq. (9) under a pooled one. The pairs (f_c, D_c) are the fixed point of Eq. (8) per channel, solved jointly because $\boldsymbol{\Psi}$ depends on all of them through μ, with one change: S_c and I_c are summed over $\operatorname{dilate}(\arg\max_k p_k = c,\ r{=}3)$, while P_c and N remain whole-patch. Without this restriction, the small probability a model assigns to a thin class inside a thicker one's territory, amplified by $D_c/2Cf_c$, creates label mass for it, and the thin channel's f_c came out $7.7\times$ too large. A channel the patch does not contain then has an empty support, so $S_c = I_c = 0$, the estimator returns $f_c = P_c/2N$ with D_c at its floor, and Eq. (19) suppresses the channel; detecting and excluding such channels instead is worse.

Results. Table 6 applies the link to the three tasks with a native multi-class version. Without labels, it lowers NLL on every column, by 14 to 58%, and the classwise ECE on every column but KiTS `3d_fullres`, where NLL falls by 20% and ECE rises slightly ($.0667 \to .0681$). A rescaling chained after the map gives the best NLL on five of the six columns. The oracle row, the same map at label-measured (f_c, D_c), is better on five of the six columns, by up to 0.170 of NLL on liver `3d_fullres`. This gap lies in D_c: f_c is recovered as at $C = 1$, whereas $I_c = \sum_v p_{cv} Y_{cv}$, estimated by $\sum_v p_{cv} \Psi_c(\mathbf{p}_v)$, is optimistic where the model is confidently wrong, which the output alone cannot tell from confidently right, and at $C > 1$ the map depends on D_c through μ.

method	Panc. l (C=2) NLL	Panc. l (C=2) ECE	Panc. f (C=2) NLL	Panc. f (C=2) ECE	Liver l (C=2) NLL	Liver l (C=2) ECE	Liver f (C=2) NLL	Liver f (C=2) ECE	KiTS l (C=3) NLL	KiTS l (C=3) ECE	KiTS f (C=3) NLL	KiTS f (C=3) ECE
label-free: nothing is fitted (bold: best in this block)												
raw output*	1.1388	.1081	.8772	.1216	.6793	.0857	.5235	.1711	.4185	.0369	.3955	**.0667**
$\mathbf{\Psi}$ (19)*	**.5656**	**.0760**	**.4922**	**.1081**	**.2843**	**.0612**	**.4485**	**.1473**	**.1999**	**.0256**	**.3161**	.0681
oracle reference: the same map at label-measured (f_c, D_c) (excluded from the marks)												
$\mathbf{\Psi}$ (19), measured pair	.5961	.0714	.4564	.0837	.2250	.0464	.2784	.0905	.1845	.0200	.2263	.0449
fitted on held-out labels, one map per model (bold: best in this block)												
beta (one-vs-rest)	.4758	.0904	.4827	.1233	.3161	.0838	.4530	.1669	.2516	.0379	.3208	.0704
isotonic (one-vs-rest)	.4670	**.0868**	.4821	.1238	.3212	.0849	.4558	.1715	**.2507**	.0374	.3191	.0729
Dirichlet	.4713	.0907	.4736	.1206	**.3060**	**.0797**	**.4401**	**.1606**	.2539	**.0373**	**.3147**	**.0657**
matrix scaling	**.4558**	.0868	**.4691**	**.1171**	.3086	.0806	.4510	.1658	.2555	.0377	.3213	.0692
a rescaling chained after the $\mathbf{\Psi}$ *map* (bold: beats the same family unchained)												
beta $\circ$ $\mathbf{\Psi}$	**.4459**	**.0808**	**.4750**	**.1218**	**.2717**	**.0719**	**.4433**	**.1506**	**.1942**	**.0295**	**.3132**	**.0664**
isotonic $\circ$ $\mathbf{\Psi}$	**.4495**	**.0811**	**.4706**	**.1189**	**.2707**	**.0710**	**.4501**	**.1500**	**.1945**	**.0290**	**.3112**	**.0659**
Dirichlet $\circ$ $\mathbf{\Psi}$	**.4375**	**.0812**	**.4665**	**.1193**	**.2630**	**.0682**	.4421	**.1481**	**.1937**	**.0276**	**.3108**	**.0640**

Table 6: **The derived link at $C > 1$.** Pancreas and liver at $C = 2$ and KiTS23 at $C = 3$, both configurations, in the layout of Table 1. $\mathbf{\Psi}$ is Eq. (19) at the label-free pair of §J, per sample for `3d_lowres` and batch-effective for `3d_fullres`; rows marked * use no labels. The oracle row uses the labels only to set (f_c, D_c) and takes no mark. Metrics are NLL and the one-vs-rest calibration error averaged over the $C + 1$ channels, each a mean over patches. Beta and isotonic are fitted one-vs-rest and renormalized; Dirichlet (Kull et al., 2019) and matrix scaling (Guo et al., 2017) are multi-class. Every fitted row is leave-one-case-out. Bold as stated in each block header; underline marks the column's best.

K Deferred proofs

Proposition 1 is $R'(q^*) = 0$, with $R' = -\eta\, g_1 + (1-\eta)\, g_0$ by Eq. (1). Proposition 3 is Eq. (3) with CE's pulls $1/q$ and $1/(1-q)$ plus the constants of Eq. (5), multiplied through by $fq(1-q)$ using $a + b = 1/f$. The decision threshold of Sec. 3.3, the fixed point $\Psi_{f,D}(D/2) = D/2$, the midpoint slope of Eq. (15) and Proposition 7 follow from Eq. (6) by direct substitution. Proposition 5 is the triangle inequality $|\Psi_{f,D}(q) - \Psi_{f',D'}(q)| \le |m(q) - \Psi_{f,D}(q)| + |m(q) - \Psi_{f',D'}(q)|$, with $\eta = \Psi_{f,D}(q)$ on the first image and $\eta = \Psi_{f',D'}(q)$ on the second.

Proposition 2 (region pulls). By Eqs. (1) and (4), $N\,\partial L/\partial q_j = -Y_j\big[g_1^{\mathrm{p}}(q_j) + a(Y)\big] + (1 - Y_j)\big[g_0^{\mathrm{p}}(q_j) + b(Y)\big]$. Taking the expectation given x with $\mathbb{E}[Y_j\, h(Y) \mid x] = \eta_j\, \mathbb{E}[h(Y) \mid x, Y_j = 1]$ and $\mathbb{E}[(1 - Y_j)\, h(Y) \mid x] = (1 - \eta_j)\, \mathbb{E}[h(Y) \mid x, Y_j = 0]$ gives

$$N\,\frac{\partial R}{\partial q_j} = -\eta_j\big(g_1^{\mathrm{p}}(q_j) + a_j\big) + (1 - \eta_j)\big(g_0^{\mathrm{p}}(q_j) + b_j\big),$$

and setting it to zero is Eq. (2) with the stated pulls.

Proposition 4 (shape). The end slopes λ_0 and λ_1 of Sec. 5 are the limits of Eq. (6) at $q \to 0$ and $q \to 1$, and both exceed 1. For a beta map, as $q \to 0$, $\operatorname{logit}\mu = c + \alpha \log q + o(1)$, so $\mu/q \sim e^c q^{\alpha-1}$, which has a finite non-zero limit only at $\alpha = 1$. Symmetrically, as $q \to 1$, $(1-\mu)/(1-q) \sim e^{-c}(1-q)^{\beta-1}$ forces $\beta = 1$. At $\alpha = \beta = 1$, $\operatorname{logit}\mu = \operatorname{logit} q + c$, whose end slopes e^c and e^{-c} multiply to 1. Platt and temperature scaling are the beta maps with $\alpha = \beta$.

Proposition 6 and Theorem 1 (edges and criterion). By Eq. (1), $R'(q) = g_1(q)\,\big[(1-\eta)\, g_0(q)/g_1(q) - \eta\big]$. Under Assumption 1 the bracket is strictly increasing, so R' changes sign at most once, from negative to positive, and the minimizer is interior exactly when $R'(0^+) < 0 < R'(1^-)$, that is, when $\eta_{\mathrm{lo}} < \eta < \eta_{\mathrm{hi}}$. For the theorem, $g_1(0^+) = g_1^{\mathrm{p}}(0^+) + a$ and $g_0(0^+) = g_0^{\mathrm{p}}(0^+) + b$, the latter finite as the infimum of a non-decreasing function. ($\Leftarrow$) $g_1^{\mathrm{p}}(0^+) = \infty$ makes $\eta_{\mathrm{lo}} = 0$ in Eq. (12), and symmetrically $g_0^{\mathrm{p}}(1^-) = \infty$ makes $\eta_{\mathrm{hi}} = 1$. ($\Rightarrow$) If $g_1^{\mathrm{p}}(0^+)$ is finite, $g_0(0^+) \ge b > 0$ gives $\eta_{\mathrm{lo}} > 0$, and every $\eta < \eta_{\mathrm{lo}}$ has $q^* = 0$; symmetrically, a finite $g_0^{\mathrm{p}}(1^-)$ and $a > 0$ give $\eta_{\mathrm{hi}} < 1$. This is the only use of $a, b > 0$, which soft Dice satisfies whenever $D > 0$.

Proposition 8 (one root). Write $u = q(1-q)$, $\langle\cdot\rangle$ for the mean over voxels and $\bar{q} = \langle q\rangle$. By Eq. (6), $\Psi_{f,D}(q) = A + DB$ with $A = qf/(f+u)$ and $B = u/(2(f+u))$. The first equation of Eq. (8) reads $2f - \bar{q} = \langle A\rangle + D\langle B\rangle$; it makes the denominator of the second equal to $2Nf$, so the second reads $Df = \langle qA\rangle + D\langle qB\rangle$. Any solution has $\bar{q}/2 \le f \le (\bar{q}+1)/2$, since $0 \le \Psi \le 1$, and $\langle qB\rangle < \bar{q}/2 \le f$. On $f \ge \bar{q}/2$ the two equations therefore define D as two functions of f,

$$D_1(f) = \frac{2f - \bar{q} - \langle A\rangle}{\langle B\rangle}, \qquad D_2(f) = \frac{\langle qA\rangle}{f - \langle qB\rangle}, \tag{20}$$

and the solutions are their crossings. D_1 *rises.* $\partial_f\langle A\rangle = \langle qu/(f+u)^2\rangle \le \bar{q}/4f \le \frac{1}{2}$, by $(f+u)^2 \ge 4fu$, so the numerator of D_1 rises with slope at least $\frac{3}{2}$ from a negative value at $f = \bar{q}/2$. It vanishes at a single f_1, beyond which D_1, a positive increasing numerator over the positive decreasing $\langle B\rangle$, increases strictly to ∞. D_2 *falls.* D_2' has the sign of $\partial_f\langle qA\rangle\,(f - \langle qB\rangle) - \langle qA\rangle\,(1 - \partial_f\langle qB\rangle)$. Here $\partial_f\langle qA\rangle\,(f - \langle qB\rangle) < \langle q^2 uf/(f+u)^2\rangle < \langle q^2 f/(f+u)\rangle = \langle qA\rangle$ and $\partial_f\langle qB\rangle \le 0$, so D_2 decreases strictly, towards 0. A solution has $D = D_2(f) > 0$, hence $f > f_1$, where $D_1 - D_2$ increases strictly from a negative value at f_1 to positive values, so it vanishes exactly once. $D \le 1$ because $q + \Psi - 2q\Psi = q(1-\Psi) + \Psi(1-q) \ge 0$.