

Contents

It Detects More, It Scores Less: A Pre-Registered Blob Loss Refutes Its Own Precision Prediction and Pays Only as a Consensus Veto under the Official BraTS-2023 Metrics

Abstract	2
1. Introduction	2
2. Related work	3
3. Methods	4
3.1 What was frozen before the measurement	4
3.2 Data, backbone and loss	5
3.3 Evaluation: one capsule, two decoding regimes	5
3.4 Statistics	6
3.5 The noise floor	6
4. Results	6
4.1 The pre-registered primary criterion: H1 refuted in both clauses	6
4.2 Primary endpoint: a deficit that is constant in sign and real in frequency	8
4.3 Three auxiliary terms, one endpoint: only blob leaves the noise	9
4.4 Raw against clean: the post-processing is the stage that decides	10
4.5 The mechanism as counts	11
4.6 Per fold: a constant sign across five independent populations	12
4.7 Fold 0 alone: three seeds and a non-determinism control	13
4.8 Where the arm does pay: as a corroborator in an ordered veto	14
4.9 Reproducibility and controls	15
5. Discussion	15
6. Limitations	17
7. Conclusion	17
Author contributions	18
Data, code and reproducibility	18
References	18

It Detects More, It Scores Less: A Pre-Registered Blob Loss Refutes Its Own Precision Prediction and Pays Only as a Consensus Veto under the Official BraTS-2023 Metrics

Guillaume Cassez · Stanislas Larnier

Independent research

Guillaume Cassez — ORCID 0009-0007-0987-3931 · cassez.guillaume@gmail.com · guillaume-cassez.fr

Stanislas Larnier — stanislaslarnier@gmail.com · HAL stanislaslarnier

Version 1 — first complete draft, written from the corpus consolidated on 2026-09-30T14:07:45. Third term of the BRATS auxiliary-loss programme to be measured on the full five-fold cross-validation, after the signed-distance and the boundary-loss arms. Everything here is read from the consolidated corpus that **Provenance**, just below, names file by file: 16 artefacts whose md5 is recorded (§4.9), under 36/36 internal controls. This prose carries no hand-typed number: it is assembled from a template in which every value is a marker resolved from that JSON, and the assembly is deterministic — regenerating it reproduces this file byte for byte (`--verif`).

Provenance. Quantitative claims come from `scripts/paper5_collector_chiffres.py` (`analysis/blobloss_stats.json`); tables from `scripts/paper5_tables.py`, which writes `table_paper5.md` (English), `table_paper5_fr.md` (French mirror), `.tex` and `.json` into `tables/`, the last carrying the JSON path of every value written; figures from `scripts/paper5_figures.py` into `figures/`, with the manifest `EMPREINTES.json`. Metrics are computed by the official BraTS-2023

capsule (external/BraTS-2023-Metrics) (Saluja et al. 2023), never by a local reimplementation. The complete official metric set is reported; the pre-registered primary criterion is separated from the primary endpoint and both from exploratory analyses; the negative result is stated in the title, the abstract and the conclusion.

Scope caveat (load-bearing). Three things this paper is *not*. It is not a study of the blob-loss weight: only $\beta = 0.5$ was ever trained, and the $\beta = 1.0$ arm announced in §4 of the pre-registration was never launched, so sensitivity to that weight is **unmeasured** (§6). It is not a study of a learned combination: the only operators examined are the twelve deterministic ordered two-expert vetoes. Nothing is learned on cached expert probabilities, and a learned mixer that this arm could feed as one expert among others — a patch-wise mixture-of-experts is the obvious candidate — is future work (§5), not a result of this paper. And it is not a proxy-metric study: the lesion counts produced by the removed post-hoc harness (outputs/pruning/lesions_*, gt_lesion_sizes.json, fit 180 / reserved 60) are excluded from every table and figure, because mixing two evaluation scales is how a programme loses track of its own numbers.

Abstract

Instance-wise losses are advertised as the remedy for small-lesion under-segmentation: they score each lesion separately, so a missed small instance costs as much as a missed large one. We pre-registered a test of one of them — blob loss (Kofler et al. 2023) — on BraTS-2023 adult glioma *before reading any lesion-wise number*, predicting from the source paper’s multiple-sclerosis cohort that it would move a MedNeXt-B / nnU-Net baseline (Roy et al. 2023 ; Isensee et al. 2021) toward **precision**: no drop in ground-truth lesions at Dice = 0, and fewer lesion false positives. Both clauses are refuted, in opposite directions. On 1195 held-out cases pooled over five disjoint folds and scored by the official BraTS-2023 capsule, the blob arm misses 129 ground-truth lesions against 190 for the baseline (significant in 3/3 regions after Holm) and produces 209 lesion false positives against 1418 (3/3 regions): it **buys recall and pays in precision**, the exact inverse of the prediction. The official post-processing then absorbs 93.2 % of the raw Dice gap and 96.2 % of the raw HD95 gap, leaving -0.00376 lesion-wise Dice (95% CI [-0.00648, -0.00103], Holm $p = 8.0e-14$) — 1.25 times the pre-specified smallest effect of interest (0.003) and negative on 5 folds out of 5. Of the three auxiliary terms we have measured on this backbone, blob is the only one whose confidence interval excludes zero, and it does so behind: signed-distance regression gives -0.00060 (Holm $p = 1$) and boundary loss -0.00116 (Holm $p = 0.9547$), both indistinguishable from the baseline. The arm’s single positive delta is as a **corroborator**: the ordered veto cc_B_G ranks 1 of 12 at +0.001137, below the effect of interest and not significant after Holm ($p = 0.156$), and reversing the order drops it to rank 11 (-0.002306). The finding is not that the loss fails; it is that a directional prediction made in blind was contradicted by measurement, and that the sign of an instance-wise loss is a property of the dataset’s instance-size distribution, not of the loss.

1. Introduction

Segmentation losses that average over a volume treat a patient’s lesions as one mass. When a scan carries a few large components and many small ones, the small ones contribute almost nothing to the gradient, and a model can lose every one of them while its Dice barely moves (Kofler et al. 2023 ; Kofler et al. 2023). Instance-wise losses remove that averaging: blob loss scores each ground-truth component against the prediction restricted to it, so every lesion carries the same weight in the objective regardless of its volume. The metric they mimic is exactly the one the BraTS-2023 protocol uses to rank submissions — lesion-wise Dice and HD95, computed per connected component (Saluja et al. 2023). The question this paper asks is therefore not exotic: **if the loss is aligned with the ranking metric, does the model move on that metric, and in which direction?**

Two of us have already asked the question twice on this backbone. Paper 1 added a signed-distance-transform regression head (Ma et al. 2020) and found a recall-oriented regulariser that does not beat the baseline; Paper 2 added a boundary loss (Kervadec et al. 2019 ; Kervadec et al. 2021) and found its mirror image on the precision side, also without beating the baseline. Those two results define an axis — recall against precision — and leave a third term untested. Blob loss is the natural third point, and the most interesting one, because it is the only term of the three whose construction is *instance-wise* rather than *boundary-wise*: it does not sharpen contours, it re-weights lesions.

Our own measured deficit sits on the recall side of that axis, which is what motivated the work: at the time the pre-registration was written, the baseline still missed a set of ground-truth lesions entirely, and those missed lesions were

orders of magnitude smaller than the detected ones. The prediction we froze was nevertheless the **opposite** of that intuition, because the primary source points the other way. In Kofler et al.’s multiple-sclerosis cohort — the setting closest to “many small lesions” — instance sensitivity *falls* (0.854 to 0.822) while precision rises (0.711 to 0.828) — figures our pre-registration recorded from a full reading of the source and that we have not re-verified online since — and their §6.9 gives the mechanism: the blob term puts false positives in the denominator of *every* instance, so spurious components are penalised proportionally to how many instances the case has. On their other datasets (LiTS, SHANEL, DeepMACT) instance sensitivity rises instead. The sign is not fixed by the loss.

So we pre-registered. docs/blob_loss_preregistration.md was written on 2026-08-06 at 12:10, while the first run was at epoch 11 of 300 and **before any lesion-wise number had been read**. It freezes the question, a directional hypothesis H1 (a shift toward precision: lesions at Dice = 0 not lower, lesion false positives lower), a null hypothesis H2 (lesion-wise Dice neutral, $|\Delta| < 0.003$), the primary criterion (the per-patient paired count of ground-truth lesions at Dice = 0, signed Wilcoxon with Holm correction over the three regions), and the guard-rails (official S1 decoding, paired comparison against the baseline at the same fold and seed, at least three seeds whenever the observed gap falls under 0.005). A framing written after reading the results would have no evidentiary value; this one is dated before them, and §4.1 reproduces it as it stands.

What the measurement says is short to state and uncomfortable to have predicted. **H1 is refuted in both of its clauses, and in opposite directions:** in the raw decoding regime the blob arm misses 129 ground-truth lesions against 190 — an improvement, significant in 3/3 regions after Holm — while producing 2029 lesion false positives against 1418, also significant in 3/3 regions. **H2 is refuted too, on the unfavourable side:** the pooled lesion-wise Dice delta is -0.00376 with a confidence interval that excludes zero. The loss does what its construction says — it recovers small instances — and pays for it with spurious ones, on a dataset where gliomas form a few large components surrounded by small satellites. The official component cleaning then removes most of both movements, and what survives is a small deficit.

Contributions:

1. **A pre-registered negative result on the official metric.** We report the primary criterion exactly as it was frozen, in both decoding regimes, together with the endpoint the programme uses for ranking (Table 2, Fig. 1). The prediction was directional and made in blind; it is contradicted; we publish the contradiction rather than the deficit, because the deficit is small and the contradiction is not.
2. **A measurement of what decides the verdict.** The official post-processing absorbs 93.2 % of the raw Dice gap and 96.2 % of the raw HD95 gap (Table 5, Fig. 2). Reporting the raw regime would have announced a collapse (-0.0554 Dice, +20.6579 mm) that the official protocol does not measure; reporting only the clean regime would have hidden the mechanism. Both are shown, and the endpoint is declared once.
3. **The mechanism stated as counts, not as a ratio.** On 4007 ground-truth lesions the blob arm detects 3878 against 3817 for the baseline, and leaves 2029 unmatched predicted components against 1418 — a 43 % rise (Tables 6a, 6b). After official cleaning the detection advantage is gone (306 missed against 310) while the false-positive gap narrows without closing (72 against 57).
4. **A bounded positive.** The arm’s only delta above the baseline is as a veto member: cc_B_G ranks 1 of 12 ordered two-expert vetoes at +0.001137 (Holm $p = 0.156$), under the effect of interest, and the veto is asymmetric — reversed it falls to rank 11 at -0.002306 (Table 9). What survives the veto is the baseline’s component, not the blob arm’s: a corroborator, not a segmenter.
5. **A noise floor that forbids the easy reading.** Three seeds at fold 0 span 0.00378 on the clean lesion-wise Dice — wider than the effect of interest and wider than the fold-0 gap itself (-0.00081) — and a second run of the baseline at the *identical* seed moves +0.00134 (Table 8, Fig. 4). No conclusion in this paper rests on a single fold.

2. Related work

Instance-wise losses. Blob loss (Kofler et al. 2023) isolates each ground-truth component by masking the prediction to it, keeping false positives in play, and averages the base metric over components; the same authors’ Panoptica framework (Kofler et al. 2023) provides the evaluation-side counterpart. The construction is what makes its behaviour dataset-dependent: because every instance’s score is divided by a term that includes all false positives, the penalty on spurious components scales with instance multiplicity. Two independent replications frame our result. Bouteille et al. (Bouteille et al. 2026) benchmark DiceCE, BlobDiceCE and their CC-DiceCE inside nnU-Net on five cohorts including BraTS 2020 (369 cases); on BraTS the blob arm holds volumetric Dice (0.9168 against 0.9171) but loses instance-level precision

(0.7159 against 0.7708) and instance-wise F1 (0.4789 against 0.4933), and the authors note that gliomas form few, very large components with small annotation artefacts nearby — the same structure our counts show. Kundu et al. (Kundu et al. 2026) extend blob loss to the multi-class setting by one-vs-rest decomposition and report a **positive** result on BraTS-METS 2025 (260 test cases): foreground Dice 0.615 against a 0.59 ± 0.27 baseline, resection-cavity Dice raised from 0.150 to 0.221, and the best panoptic quality at threshold 0.5 (0.40 ± 0.24 against 0.38 ± 0.25). Read together with ours, the three studies cover three instance-size rega — many tiny lesions (metastases, positive), few large components with small satellites (glioma, precision-negative in both independent measurements), and small multiple sclerosis lesions (precision-positive in the source paper). The sign follows the regime, not the loss.

Auxiliary terms on a strong backbone. Signed-distance regression (Ma et al. 2020 ; Xue et al. 2020 ; Dangi et al. 2019) and boundary loss (Kervadec et al. 2019 ; Kervadec et al. 2021) are the two terms Papers 1 and 2 of this programme measured on the same MedNeXt-B / nnU-Net v2 backbone; both are re-scored here on the identical population and the identical Holm family as comparators (Table 4), not re-trained. Region-based losses and ensembles for BraTS are the standard alternatives (Isensee et al. 2020 ; Abe et al. 2022 ; Theisen et al. 2023), and consensus operators built from independently-trained models are the training-free family this paper’s veto belongs to (Warfield et al. 2004 ; Wolpert 1992 ; Ding et al. 2019).

Backbone, data and metrics. nnU-Net (Isensee et al. 2021) with the MedNeXt-B block (Roy et al. 2023); Dice and generalised Dice (Milletari et al. 2016 ; Sudre et al. 2017); BraTS glioma (Menze et al. 2015 ; Bakas et al. 2017 ; Baid et al. 2021); the official BraTS-2023 lesion-wise metrics and their surface-distance companion (Saluja et al. 2023 ; Moawad et al. 2023 ; Nikolov et al. 2021).

Why the protocol is stated before the result. Ranking on a single fold, on a single seed, or on a metric that is not the challenge’s is the documented failure mode of this literature (Maier-Hein et al. 2018 ; Maier-Hein et al. 2024 ; Reinke et al. 2024 ; Christodoulou et al. 2025). Seed variance alone can exceed a published effect (Picard 2021 ; Bouthillier et al. 2021 ; Åkesson et al. 2024), cross-validation folds are not independent samples (Nadeau et al. 2003), and a null hypothesis test cannot establish an absence of effect without an equivalence bound (Berger 1982). Negative results are worth publishing precisely because the literature under-reports them (Varoquaux et al. 2022). Every one of those points is operationalised below: a pre-specified smallest effect of interest, Holm correction inside *declared* families, three seeds plus a same-seed re-run control, and a pooled cross-validation as the endpoint rather than any single fold.

3. Methods

3.1 What was frozen before the measurement

The pre-registration (docs/blob_loss_preregistration.md, 2026-08-06 12:10) fixes four things that this paper does not get to choose afterwards:

- **H1 (directional).** The blob term moves the model toward *precision*: the number of ground-truth lesions at Dice = 0 does **not** fall, and lesion false positives fall. The prediction was deliberately written against our own intuition, on the strength of the source paper’s multiple-sclerosis cohort.
- **H2 (null).** Global lesion-wise Dice is neutral: $|\Delta| < \text{SESOI} = 0.003$.
- **Primary criterion.** The per-patient paired count of ground-truth lesions at Dice = 0, signed Wilcoxon, Holm over the three regions (WT, TC, ET). In the official BraTS-2023 scorer a ground-truth lesion at Dice = 0 is an unmatched ground-truth lesion (`num_fn`); that definitional equivalence is not assumed, it is asserted by the consolidation against the pooled counts in both populations and both regimes (Table 10).
- **Guard-rails.** Official S1 decoding; paired comparison against the baseline at the same fold and seed; **at least three seeds** whenever the observed gap is under 0.005, because the measured seed-to-seed noise on a free head is $\text{sd} = 0.005$ on the voxel-wise scale, above the effect of interest; and no conclusion drawn from a short read-out.

Two of these were honoured and one was not fully. Three seeds were trained at fold 0, as required. The $\beta = 1.0$ arm listed in §4 of the pre-registration was never launched, so the weight sensitivity of every claim below is unknown (§6). The pre-registration also declared that this work would be a **standalone paper** in the auxiliary-loss family rather than a section appended to another study; that is what this document is.

3.2 Data, backbone and loss

BraTS 2023 adult glioma, nnU-Net dataset of 1196 cases after the 55 documented exclusions, split into five disjoint folds; held-out validation covers 240 cases at fold 0 and 239 at each of folds 1 to 4. One case, BraTS-GLI-00017-001, carries an empty ground truth, for which lesion-wise Dice is undefined; it is removed from the pooled population, giving **n = 1195**, the population on which every other arm of this programme is scored. The full n = 1196 population is computed as well and changes no conclusion.

The backbone is MedNeXt-B under nnU-Net v2 with the nnUNetPlans_96GB_mednext plans, trained 300 epochs on a single NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition. The arm differs from the reference by one term:

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}} + \beta \cdot \mathcal{L}_{\text{blob}}, \quad \beta = 0.5, \quad \lambda_{\text{dist}} = 0.0$$

with $\beta = 0.5$ reproducing Kofler et al.’s 2:1 global-to-blob ratio, and the signed-distance term of Paper 1 switched off. Both parameters are read from the `debug.json` of **every** run rather than from documentation, and the epoch count from the training log — the `debug.json` of fold 0 carries a stale `current_epoch` of 15 where the log shows 300 completed epochs, which is exactly why the two sources are not interchangeable. Seven runs were trained: five folds at seed 42 and two extra seeds (1337, 2024) at fold 0. The campaign closed 2026-08-20T07:20:36Z with 7 of 7 runs completed and 0 failures (Table 1).

A note on the number that attests convergence. nnU-Net writes a *Mean Validation Dice* at the end of training: 0.90622 averaged over the five folds, with a spread of 0.00184 across the three seeds at fold 0. That number is **voxel-wise**, computed with test-time-augmentation mirroring and without component cleaning. It is not the official BraTS metric, it is not comparable to anything in Tables 2 to 9, and it appears in this paper for one purpose only: to attest that every run trained to budget.

Table 1 — What was trained (campaign attestation)

Mean Validation Dice is the **voxel-wise** number nnU-Net writes at the end of training, with TTA mirroring and no connected-component cleaning. It is **not** the official BraTS metric and it is reported here only to attest that every run converged for the full budget. All official results are in Tables 2 to 9.

Run	Epochs	beta (blob)	lambda_dist	Network / plan	Mean Validation Dice (voxel-wise, nnU-Net log)	Validation predictions
seed42_fold0 — Fold 0, seed 42	300	0.5	0.0	MedNeXt	0.906820	240
seed42_fold1 — Fold 1, seed 42	300	0.5	0.0	MedNeXt	0.902310	239
seed42_fold2 — Fold 2, seed 42	300	0.5	0.0	MedNeXt	0.908328	239
seed42_fold3 — Fold 3, seed 42	300	0.5	0.0	MedNeXt	0.910391	239
seed42_fold4 — Fold 4, seed 42	300	0.5	0.0	MedNeXt	0.903231	239
seed1337_fold0 — Fold 0, seed 1337	300	0.5	0.0	MedNeXt	0.904983	240
seed2024_fold0 — Fold 0, seed 2024	300	0.5	0.0	MedNeXt	0.905909	240

Campaign: **7 runs, 7 completed, 0 failed**, closed 2026-08-20T07:20:36Z. Mean over the 5 folds (seed 42): **0.90622**; spread of the 3 seeds at fold 0: **0.00184**.

3.3 Evaluation: one capsule, two decoding regimes

All metrics come from the official BraTS-2023 capsule (`external/BraTS-2023-Metrics`) at the **lesion-wise** scale: Dice and HD95 per region (WT, TC, ET), computed over connected components rather than over the volume. Two decoding regimes are reported side by side throughout, because the difference between them is one of this paper’s findings:

- **clean — the official regime.** Argmax decoding followed by per-region removal of connected components smaller than 1000 voxels (WT), 250 (TC) and 500 (ET). This is what the challenge scores, and it is the declared primary regime of this paper.
- **raw — no cleaning.** The same argmax prediction with no component filtering. It is reported because the blob loss is *designed* to change the component-size distribution, and a regime that deletes small components before scoring hides exactly the effect under study.

Pooled cross-validation covers 200928 scored rows ($28 \text{ arms} \times 2 \text{ regimes} \times 3 \text{ regions} \times 1196 \text{ cases}$); the fold-0 evaluation adds 7200 measures ($5 \text{ arms} \times 2 \text{ regimes} \times 3 \text{ regions} \times 240 \text{ cases}$).

3.4 Statistics

Paired Wilcoxon signed-rank per patient (Wilcoxon 1945), Holm correction inside **declared** families (Holm 1979): the expert family $\{G, D, K\} \times \{WT, TC, ET, \text{mean}\}$ for the primary endpoint; the three regions for the pre-registered primary criterion, as frozen; the twelve vetoes; and 20 tests ($5 \text{ folds} \times 4 \text{ regions}$), clean regime, lesion-wise Dice for the per-fold decomposition. Confidence intervals are 95% on the paired mean; the minimum detectable effect is $MDE = 2.802 \times \text{se}$. The smallest effect of interest is pre-specified at $SESOI = 0.003$, and the ratio $|\Delta|/SESOI$ is reported next to every headline delta so that a reader never has to reconstruct it.

Two instruments are reported together everywhere, because either one quoted alone would mislead. The paired **mean** is blunt on these data: paired sd 0.0481 for a delta of -0.00376, hence nominal MDE 0.00390, larger than the effect it is supposed to detect. The **signed-rank** test is what rejects, and what it says is not “much worse” but “worse more often than better”: of 1193 cases whose score moved, 713 went down. The publishable verdict is the conjunction — *a small deficit, constant in sign, real in frequency, of a magnitude that sits around the threshold of clinical interest* — and no single number in this paper carries it alone.

One row of the pre-registered criterion is significant after Holm while its mean-difference interval still contains zero (1 of 12 rows across both regimes). That is not a contradiction to be smoothed over: the interval describes the paired mean, the p-value the signed ranks, and on counts that are zero for nearly every case the effect lives in how often a count moves rather than in how far it moves on average. On the primary endpoint, 2 of the 4 reported rows are significant with an interval that spans zero (TC and ET); both are shown as such in Table 3.

3.5 The noise floor

A gap measured at one fold cannot be interpreted without knowing what the same training command produces twice. Two controls are trained for that:

- **three seeds** of the blob arm at fold 0 (42, 1337, 2024), the replication the pre-registration demanded;
- **a second run of the baseline at the identical seed**, which measures GPU non-determinism and nothing else — not a model difference.

They are reported in Table 8 and Fig. 4 with the fold-0 gap between them. The verdicts written by the evaluation are the ones used here verbatim: *indiscernable* (dans l'etendue des 3 graines) in the clean regime, *hors du bruit du témoin* in the raw regime.

4. Results

4.1 The pre-registered primary criterion: H1 refuted in both clauses

The criterion was frozen as the paired per-patient count of ground-truth lesions at Dice = 0, with Holm over the three regions, and H1 predicted it would **not** fall while lesion false positives fell. Table 2 and Fig. 1 report it in both decoding regimes.

In the **raw** regime both clauses fail, and they fail in opposite directions. Missed ground-truth lesions drop from 190 to 129 — significant in 3/3 regions after Holm (WT $\Delta -0.02008$, $p = 7.5e-06$; TC $\Delta -0.01590$, $p = 6.8e-05$; ET $\Delta -0.01506$, $p = 2.4e-04$), with 37 cases improved against 2 degraded and 1156 unchanged. Lesion false positives rise from 1418 to 2029, again significant in 3/3 regions (WT $+0.27280$, $p = 1.0e-22$; TC $+0.17824$, $p = 6.7e-21$; ET $+0.06025$, $p = 1.9e-06$), with 408 cases degraded against 123 improved. The loss buys recall and pays in precision — the inverse of the prediction, which was drawn from a multiple-sclerosis cohort where instance multiplicity is high and the per-instance false-positive denominator does the opposite job.

In the **clean** regime the official component filter absorbs most of the movement. Missed lesions go 310 to 306, significant in 1/3 regions (WT -0.00669 , $p = 0.0342$; TC $+0.00000$, $p = 1$; ET $+0.00335$, $p = 0.3146$), with 15 cases improved against 10 degraded and 1170 unchanged. False positives go 57 to 72, significant in 0/3 regions. What the raw regime showed as a large bidirectional shift survives the protocol as a small one — which is the finding of §4.4 in a single line.

Table 2 — The pre-registered primary criterion, tested

Source of the pre-registration: docs/blob_loss_preregistrement.md §3 (written 2026-08-06, **before any lesion-wise number was read**). Criterion as frozen there, verbatim from the French original: *num_fn — lésions GT jamais apparées (Dice = 0), apparié par patient = num_fn*, ground-truth lesions never matched (Dice = 0), paired per patient; Holm family: 3 tests (les 3 régions WT/TC/ET), conformément au pré-enregistrement = 3 tests over WT/TC/ET. Equivalence with the pooled counts asserted by Table 10.

H1 as pre-registered predicted a shift toward *precision*: the number of lesions at Dice = 0 would **not** drop, and lesion false positives would drop. **Measured: both clauses are refuted, in opposite directions.**

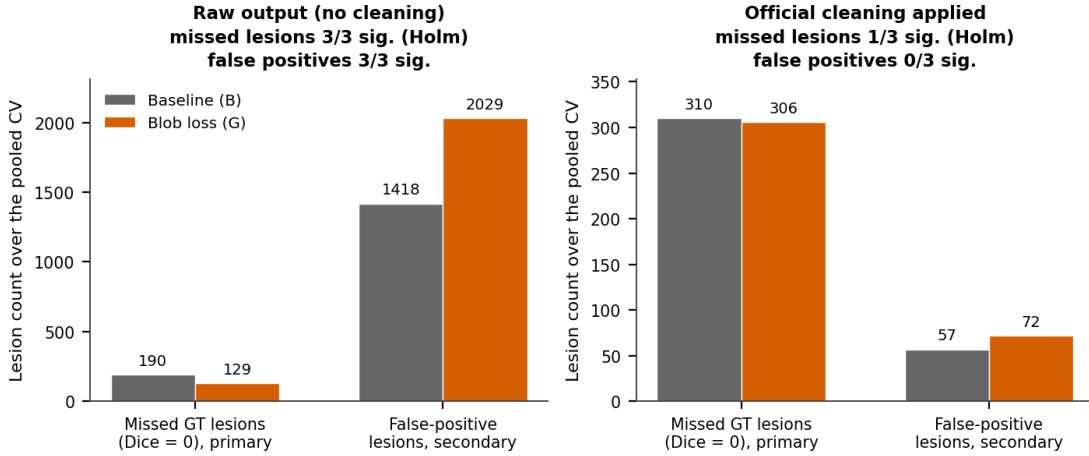
Read by group: regime, criterion and the pooled totals are written once per group, on its first and last rows.

Regime	Pre-registered criterion	Region	Delta per case (G - B)	95% CI	p (Holm)	Cases G better	Cases moved	Total G	Total B
raw	PRIMARY — missed GT lesions (Dice = 0)	WT	-0.02008	[-0.02837, -0.01179]	7.5e-06	23	23		
		TC	-0.01590	[-0.02336, -0.00843]	6.8e-05	18	18		
		ET	-0.01506	[-0.02306, -0.00707]	2.4e-04	19	21		
		pooled	—	—	—	37 lower / 2 higher / 1156 unchanged	1195	129	190
raw	Secondary — false-positive lesions	WT	+0.27280	[0.21030, 0.33531]	1.0e-22	76	352		
		TC	+0.17824	[0.13670, 0.21979]	6.7e-21	23	200		
		ET	+0.06025	[-0.02929, 0.14979]	1.9e-06	80	275		
		pooled	—	—	—	123 lower / 408 higher / 664 unchanged	1195	2029	1418
clean	PRIMARY — missed GT lesions (Dice = 0)	WT	-0.00669	[-0.01187, -0.00152]	0.0342	9	10		
		TC	+0.00000	[-0.00519, 0.00519]	1	5	10		
		ET	+0.00335	[-0.00129, 0.00798]	0.3146	2	8		
		pooled	—	—	—	15 lower / 10 higher / 1170 unchanged	1195	306	310
clean	Secondary — false-positive lesions	WT	+0.01172	[0.00189, 0.02154]	0.0629	4	19		
		TC	+0.00167	[-0.00489, 0.00824]	1	5	13		
		ET	-0.00084	[-0.00576, 0.00409]	1	5	9		
		pooled	—	—	—	11 lower / 19 higher / 1165 unchanged	1195	72	57

- **Raw regime.** Missed lesions fall from **190** (baseline) to **129** (blob), significant in 3 of 3 regions after Holm ($p = 7.5e-06, 6.8e-05, 2.4e-04$); 37 cases improve against 2 that degrade. False positives rise from **1418** to **2029**, significant in 3 of 3 regions. The loss does the **opposite** of the pre-registered prediction: it buys recall and pays in precision.
- **Clean regime.** The official cleaning erases most of it: missed lesions **310** to **306**, still significant in 1 of 3 regions (WT $p = 0.0342$); 15 cases improve against 10 that degrade, 1170 unchanged. False positives **57** to **72**, 0 of 3 regions significant after Holm.
- **H2 as pre-registered** predicted a *neutral* lesion-wise Dice ($|\text{delta}| < \text{SESOI} = 0.003$). Measured (Table 3): $\text{delta} = -0.00376$ with a CI excluding zero — **H2 is refuted too**, on the unfavourable side.
- **Two instruments, stated openly.** 1 row(s) of this table are significant after Holm while their mean-difference CI still contains zero (raw/faux_positifs/ET). That is not a contradiction: the CI describes the paired **mean**, the p-value the **signed ranks**. On counts that are zero for almost every case, the effect lives in how often the count moves, not in how far it moves on average. Both are reported so that neither can be quoted alone.

Consequence for the paper. This is not a null result dressed up as one: a directional prediction made in blind was contradicted by the measurement, and the contradiction is itself the finding — aligning the loss with the lesion-wise metric does move the model on the lesion counts, just not in the direction the source dataset (Kofler’s MS cohort) suggested, and the official post-processing then absorbs most of the movement.

The pre-registered primary criterion contradicts its own prediction
H1 expected fewer false positives and no gain in missed lesions; measured: recall bought, precision paid



Paired per patient, $n = 1195$ held-out cases (5 pooled folds). Holm over the three regions, as frozen in docs/blob_loss_preregistration.md §3 before any measurement.

Figure 1: The pre-registered primary criterion, paired per patient, in both decoding regimes. Left: ground-truth lesions missed entirely (Dice = 0), the frozen criterion; right: lesion false positives. The raw regime shows the two clauses of H1 failing in opposite directions; the clean regime shows the official component filter absorbing most of both.

H2 is refuted as well, on the unfavourable side: the pooled lesion-wise Dice delta is -0.00376 with a confidence interval excluding zero (§4.2), where H2 predicted $|\Delta| < 0.003$. Neither hypothesis survives, and the two failures are not the same kind: H1 was a directional claim about a mechanism and it is contradicted in its mechanism; H2 was an equivalence claim and it fails because the effect is real, if small.

4.2 Primary endpoint: a deficit that is constant in sign and real in frequency

Table 3 gives the declared primary endpoint — official lesion-wise Dice and HD95, mean of the three regions per patient, clean regime, pooled over five folds, $n = 1195$, paired against the baseline of the same fold and seed.

The mean-region delta is **-0.00376** (95% CI $[-0.00648, -0.00103]$, Holm $p = 8.0e-14$), 1.25 times the pre-specified SESOI, on absolute Dice 0.8776 for the baseline against 0.8738 for the blob arm. Per region the sign never changes: WT -0.00408 (CI $[-0.00764, -0.00052]$, Holm $p = 2.5e-04$), TC -0.00414 (CI $[-0.00850, 0.00022]$, $p = 1.6e-11$), ET -0.00305 (CI $[-0.00685, 0.00076]$, $p = 2.3e-10$). Lesion-wise HD95 moves the same way without reaching significance on the mean ($+0.7772$ mm, Holm $p = 1$), with ET the only region significant ($+0.6362$ mm, $p = 0.0283$) — Dice and HD95 do not have to agree, and here they agree in sign but not in strength.

Read with both instruments: the interval excludes zero, so the deficit is real; 480 of 1195 cases are better under the blob arm, and of the 1193 cases whose score moved at all, 713 moved down; but the paired mean’s own MDE (0.00390) exceeds the effect it detects, so the *magnitude* remains uncertain between “negligible” and “twice the SESOI”. The honest statement is the one in §3.4, and it is the one the abstract carries.

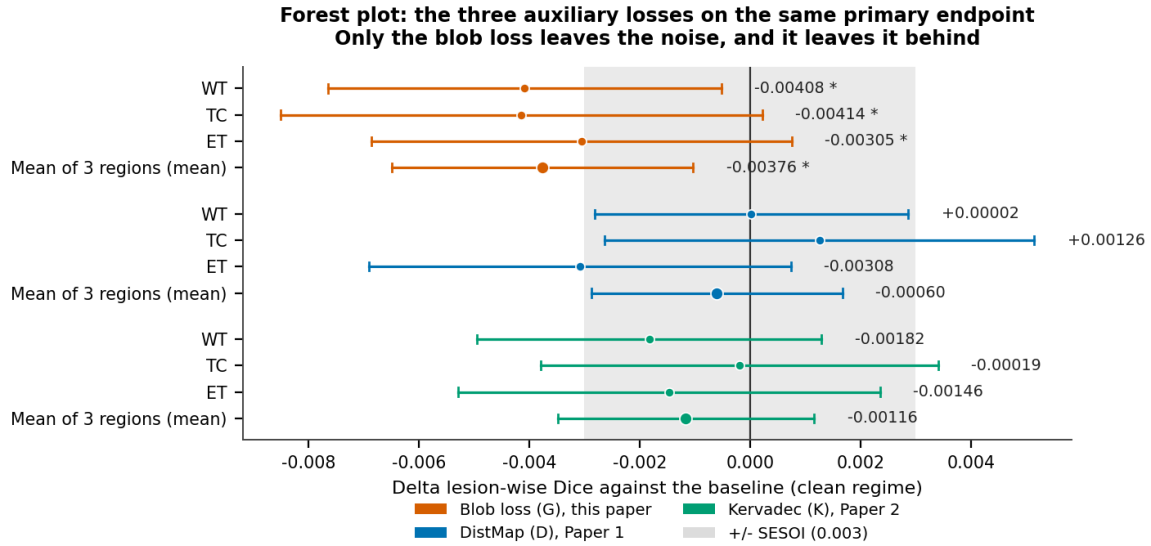
Read-out. The mean-region deficit is -0.00376 (CI $[-0.00648, -0.00103]$, Holm $p = 8.0e-14$): the interval **excludes zero**, so the deficit is real in frequency, and it is 1.25x the pre-specified SESOI (0.003). The 95% CI excludes zero for 2 of the 4 reported rows (WT, Mean of 3 regions); the other intervals span zero and are reported as such. But the paired mean is a

Table 3 — Primary endpoint: pooled 5-fold CV, official cleaning

Population **n** = **1195** held-out cases (5 folds, the degenerate case BraTS-GLI-00017-001 with an empty ground truth removed — the same population on which every other arm of this programme is scored). Official BraTS-2023 **lesion-wise** metrics, **clean** regime (per-region connected-component cleaning at 1000/250/500 voxels). Paired Wilcoxon signed-rank per patient, Holm correction over the family {G, D, K} x {WT, TC, ET, mean}. Arm G = blob loss (beta = 0.5), reference B = MedNeXt-B baseline, same plans, same fold, same seed.

Region	Dice B	Dice G	Delta Dice	95% CI	p (Holm)	Cases where G wins	MDE	Delta HD95 (mm)	95% CI	p (Holm)
WT	0.8932	0.8892	-0.00408	[-0.00764, -0.00052]	2.5e-04	539/1195	0.00509	+0.704	[-0.743, 2.151]	1
TC	0.8916	0.8875	-0.00414	[-0.00850, 0.00022]	1.6e-11	480/1195	0.00623	+0.991	[-1.037, 3.020]	1
ET	0.8479	0.8448	-0.00305	[-0.00685, 0.00076]	2.3e-10	474/1195	0.00544	+0.636	[-1.227, 2.499]	0.0283
Mean of 3 regions	0.8776	0.8738	-0.00376	[-0.00648, -0.00103]	8.0e-14	480/1195	0.00390	+0.777	[-0.457, 2.011]	1

blunt instrument here: paired sd 0.0481 for a delta of 0.0038, hence nominal MDE 0.00390 > |delta|. What the rank test says is not “much worse” but “worse more often than better”: 713 of 1193 moved cases went down. Publishable verdict: **a small, constant, real-in-frequency deficit whose magnitude stays around the threshold of clinical interest.**



* p < 0.05 after Holm. (mean) = mean of the three regions, the pre-specified endpoint.
Pooled 5-fold CV, n = 1195 held-out cases, paired Wilcoxon signed-rank, Holm over {G, D, K} x {WT, TC, ET, mean}. Whiskers are 95 % CI on the paired mean.

Figure 2: The forest plot of the primary endpoint: the three auxiliary terms against the baseline on lesion-wise Dice, with 95% confidence intervals, Holm-corrected p-values and the pre-specified SESOI drawn as a band. The blob arm is the only one of the three whose interval excludes zero, and it excludes it on the wrong side.

4.3 Three auxiliary terms, one endpoint: only blob leaves the noise

Table 4 places the two other terms of the programme on the identical population, regime and Holm family. Signed-distance regression gives -0.00060 (CI [-0.00287, 0.00167], Holm p = 1) and boundary loss -0.00116 (CI [-0.00348, 0.00116], p = 0.9547): neither interval excludes zero, and neither is distinguishable from the baseline. Blob gives

-0.00376 with an interval that does.

This is the sentence that must not be read as a ranking of merit. Within the single Holm family of 12 tests — 3 arms by 4 aggregations — the term that leaves the noise is the term with a consistent *sign*, not the term with a large effect: the two null arms are null in the strict sense of an interval straddling zero, and blob is negative in the strict sense of an interval entirely below it. All three sit under or around the SESOI in magnitude, and none is a candidate for deployment. What the comparison does establish is that “aligned with the lesion-wise metric” is not the same property as “helpful on the lesion-wise metric” — the one term built to mimic the ranking metric is also the only one measurably worse under it.

Table 4 — The three auxiliary losses on the same primary endpoint

Same population, same regime, same Holm family. DistMap (D, Paper 1) and Kervadec (K, Paper 2) are shown as comparators: neither leaves the noise. The blob loss is the **only** one of the three whose confidence interval excludes zero — and it does so on the wrong side.

Auxiliary loss	Delta Dice (mean)	95% CI	p (Holm)	Delta HD95 (mm)	p (Holm)	Cases won / n	Verdict vs SESOI
Blob loss (G)	-0.00376	[-0.00648, -0.00103]	8.0e-14	+0.777	1	480/1195	CI excludes zero - the only one of the three to leave the noise, and it leaves it behind
DistMap (D)	-0.00060	[-0.00287, 0.00167]	1	-0.079	0.1723	612/1195	CI spans zero - indistinguishable from the baseline
Kervadec (K)	-0.00116	[-0.00348, 0.00116]	0.9547	+0.464	1	643/1195	CI spans zero - indistinguishable from the baseline

4.4 Raw against clean: the post-processing is the stage that decides

Table 5 and Fig. 3 put the two decoding regimes side by side. In the raw regime the blob arm is -0.0554 on lesion-wise Dice (Holm $p = 1.3\text{e-}46$) and +20.6579 mm on lesion-wise HD95 ($p = 1.7\text{e-}27$). In the clean regime those become -0.00376 ($p = 8.0\text{e-}14$) and +0.7772 mm ($p = 1$). The official component filter therefore absorbs **93.2 %** of the Dice gap and **96.2 %** of the HD95 gap, and the absorption is uniform across regions — 92.4 % to 93.8 % of the Dice gap and 95.1 % to 97.1 % of the HD95 gap, the per-region deltas going from -0.0663 to -0.00408 on WT, -0.0543 to -0.00414 on TC and -0.0457 to -0.00305 on ET.

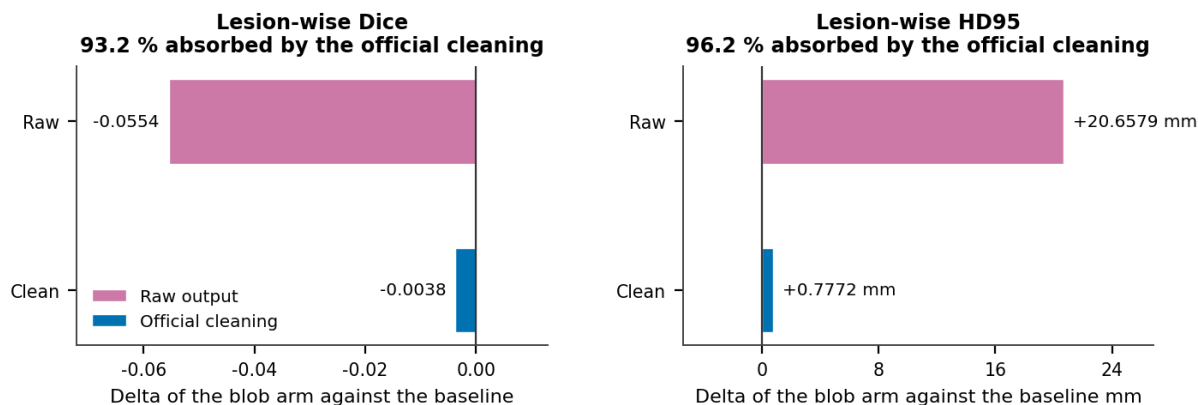
Two consequences follow, and they pull in different directions. Methodologically, a paper that reported only the raw regime would announce a collapse the challenge protocol does not measure, and a paper that reported only the clean regime would conceal the fact that its arm changes the component-size distribution enough for a filter to matter. Both regimes are therefore in every table that could carry them, with the absorbed share computed rather than asserted. Substantively, the absorption *is* the mechanism: the blob term adds small components — some of them ground-truth lesions recovered, most of them not — and the official filter removes small components indiscriminately. What survives is the residue, and the residue is a small deficit.

Table 5 — Raw vs clean: what the official post-processing absorbs

The **raw** regime applies no connected-component cleaning; **clean** applies the official per-region thresholds. The blob loss is designed to recover small instances, and in the raw regime it visibly does what it advertises: it detects more lesions. The official decoding then removes most of what it added, and the surviving signal is a small deficit.

Metric	Region	Delta raw	p (Holm) raw	Delta clean	p (Holm) clean	Share absorbed by the official cleaning
LW Dice	WT	-0.0663	1.9e-24	-0.0041	2.5e-04	93.8 %
LW Dice	TC	-0.0543	1.9e-26	-0.0041	1.6e-11	92.4 %
LW Dice	ET	-0.0457	1.2e-21	-0.0030	2.3e-10	93.3 %
LW Dice	Mean of 3 regions	-0.0554	1.3e-46	-0.0038	8.0e-14	93.2 %
LW HD95	WT	+24.5394	2.4e-13	+0.7041	1	97.1 %
LW HD95	TC	+20.0421	1.8e-12	+0.9913	1	95.1 %
LW HD95	ET	+17.3922	5.7e-11	+0.6362	0.0283	96.3 %
LW HD95	Mean of 3 regions	+20.6579	1.7e-27	+0.7772	1	96.2 %

**The official post-processing is the stage that decides
Reporting the raw regime would announce a collapse the official protocol does not measure**



Pooled 5-fold CV, $n = 1195$, mean of the three regions per patient. Connected-component cleaning per region at 1000 (WT) / 250 (TC) / 500 (ET) voxels.

Figure 3: What the official post-processing absorbs. Each pair of bars is the same arm-metric gap before and after component cleaning; the annotated share is computed from the two, not quoted. Publishing in the raw regime would have announced a collapse; publishing only in the clean regime would have hidden the mechanism.

4.5 The mechanism as counts

Tables 6a and 6b state the mechanism without ratios. On 4007 ground-truth lesions in the pooled cross-validation, the **raw** regime gives the blob arm 3878 detections against 3817 for the baseline — a detection rate of 0.9678 against 0.9526 — at the cost of 2029 unmatched predicted components against 1418, a 43 % increase. That is the whole story of the loss in one line: it finds more of what is there and invents more of what is not.

After official cleaning the detection advantage is essentially gone — 306 missed lesions against 310, detection rates 0.9236 against 0.9226 — while the false-positive gap narrows without closing (72 against 57). For comparison, the two other auxiliary arms behave as their papers describe: signed-distance regression recovers part of the recall at a moderate false-positive cost, and boundary loss is the only one of the three to *reduce* false positives in the raw regime.

The veto row of the same tables anticipates §4.8: cc_B_G keeps the baseline’s detection rate (0.9226) and removes false positives (47 against 57). That is the one place in this paper where the blob arm’s contribution is positive, and it is a contribution of *agreement*, not of segmentation.

Table 6a — Lesion counting, raw regime (no cleaning)

Pooled over the 5 folds. This is the mechanism, stated as counts rather than as a ratio: the blob arm finds **more** ground-truth lesions than the baseline and pays for it with **more** unmatched predicted components.

Arm	GT lesions	Detected	Missed	Detection rate	Unmatched predictions (FP)	FP per case
Baseline (B)	4007	3817	190	0.9526	1418	1.1866
Blob loss (G)	4007	3878	129	0.9678	2029	1.6979
DistMap (D)	4007	3839	168	0.9581	1577	1.3197
Kervadec (K)	4007	3827	180	0.9551	1311	1.0971
B vetoed by G	4007	3815	192	0.9521	903	0.7556
G vetoed by B	4007	3815	192	0.9521	886	0.7414
Majority vote (2 of 3)	4007	3824	183	0.9543	1190	0.9958

Table 6b — Lesion counting, clean regime (official cleaning)

After the official cleaning the detection advantage is gone (the two arms miss almost the same number of ground-truth lesions) while the false-positive gap narrows but does not close. The veto `cc_B_G` keeps the baseline detection rate and removes false positives.

Arm	GT lesions	Detected	Missed	Detection rate	Unmatched predictions (FP)	FP per case
Baseline (B)	4007	3697	310	0.9226	57	0.0477
Blob loss (G)	4007	3701	306	0.9236	72	0.0603
DistMap (D)	4007	3701	306	0.9236	54	0.0452
Kervadec (K)	4007	3690	317	0.9209	58	0.0485
B vetoed by G	4007	3697	310	0.9226	47	0.0393
G vetoed by B	4007	3698	309	0.9229	57	0.0477
Majority vote (2 of 3)	4007	3685	322	0.9196	52	0.0435

4.6 Per fold: a constant sign across five independent populations

Table 7 decomposes the primary endpoint fold by fold, each fold scored against the baseline of the same fold, Holm over 20 tests (5 folds $\times$ 4 regions), clean regime, lesion-wise Dice. The clean deltas are -0.00216, -0.00561, -0.00425, -0.00301 and -0.00375 on 239 cases each: **0 of 5 folds positive**, 5 negative, spread 0.00346, mean of the five -0.00376. The raw regime is unanimous as well (0 folds positive), and its spread is 4.65 times the clean one: 0.01607 against 0.00346.

The per-fold intervals mostly contain zero, and only 1 fold has a delta smaller in absolute value than the SESOI — the sign is constant while the per-fold resolution is not. That is precisely the situation the pre-registration’s guard-rail anticipated: the claim rests on the *pooled* paired test and on the constancy of the sign across five disjoint patient populations, never on a significant per-fold test. It is also why the paper does not say “the blob loss is significantly worse on every fold”: it does not have the resolution to say that, and saying it would be the over-claim the SESOI exists to prevent.

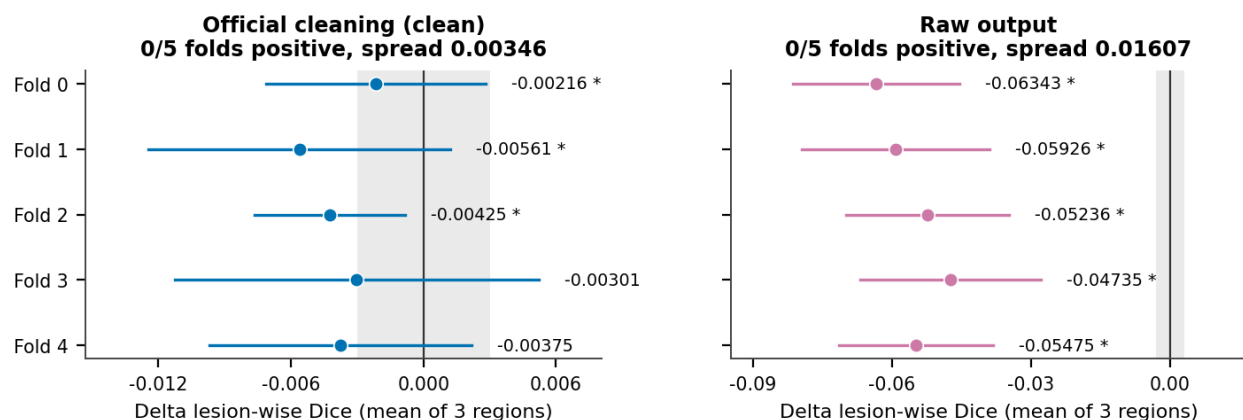
Table 7 — Per-fold decomposition of the primary endpoint

Each fold is scored separately against the baseline of the same fold, then pooled. Holm family: 20 tests (5 folds $\times$ 4 regions) per regime. The per-fold sign is constant, which is what makes the pooled interval interpretable.

Fold	n	Delta Dice clean	95% CI	p (Holm)	Delta Dice raw	p (Holm) raw
0	239	-0.00216	[-0.00714, 0.00283]	2.1e-05	-0.06343	1.5e-13
1	239	-0.00561	[-0.01245, 0.00123]	3.1e-05	-0.05926	1.0e-11
2	239	-0.00425	[-0.00765, -0.00084]	4.6e-08	-0.05236	3.7e-10
3	239	-0.00301	[-0.01125, 0.00522]	0.1295	-0.04735	8.7e-07
4	239	-0.00375	[-0.00968, 0.00218]	1	-0.05475	8.2e-06

Clean regime: **0/5 folds positive**, 5 negative, spread 0.00346, mean of the 5 folds -0.00376, 1 fold(s) below the SESOI in absolute value. Raw regime: 0/5 folds positive, spread 0.01607.

Five independent folds, one constant sign
The pooled interval of the primary endpoint is interpretable because no fold disagrees



Each fold is scored against the baseline of the SAME fold. Holm family: 20 tests (5 folds $\times$ 4 regions) per regime. * $p < 0.05$ after Holm.

Figure 4: Per-fold decomposition of the primary endpoint in both decoding regimes, with the SESOI band. The sign is constant across the five disjoint populations; the per-fold intervals are not resolvable, which is why the conclusion is pooled.

4.7 Fold 0 alone: three seeds and a non-determinism control

Table 8 is the reason no conclusion in this paper rests on one fold. At fold 0, in the clean regime, the blob arm at seed 42 sits -0.00081 from the baseline (CI [-0.00658, 0.00495], MDE 0.00824) — while a **second run of the baseline at the identical seed** sits +0.00134. The three blob seeds give -0.00081, -0.00185, +0.00194, spanning 0.00378 around a mean lesion-wise Dice of 0.88140: a spread 1.26 times the SESOI and larger than the fold-0 gap it is meant to interpret. The evaluation’s own verdict is indiscernable (dans l’etendue des 3 graines), and this paper adopts it without softening.

In the raw regime the same fold is unambiguous: -0.05130 for the blob arm against +0.00390 for the control, verdict hors du bruit du témoin. The two regimes disagree about whether fold 0 shows an effect, and the disagreement is itself informative — the effect that is large enough to clear a single fold’s noise is the one the official protocol then deletes.

Table 8 — Fold 0 alone: three seeds and a non-determinism control

Fold 0 only, $n = 239$ after removing the degenerate case. The control is a **second run of the baseline at the identical seed**: it measures GPU non-determinism, not a model difference. The three blob seeds are the pre-specified replication.

Arm	Delta Dice clean	95% CI	p (Holm)	Delta Dice raw	LW Dice of the arm (clean)
Baseline, 2nd run, identical seed (control)	+0.00134	[-0.00223, 0.00492]	0.3225	+0.00390	0.8830
Blob loss, seed 42	-0.00081	[-0.00658, 0.00495]	4.0e-05	-0.05130	0.8808
Blob loss, seed 1337	-0.00185	[-0.00827, 0.00457]	8.8e-05	-0.04448	0.8798
Blob loss, seed 2024	+0.00194	[-0.00278, 0.00666]	2.5e-04	-0.05227	0.8836

Clean regime verdict: **indistinguishable (within the 3-seed spread)** (blob -0.00081, control +0.00134, three seeds -0.00081, -0.00185, +0.00194). Raw regime verdict: **outside the control’s noise** (blob -0.05130, control +0.00390). Seed spread on the clean lesion-wise Dice: **0.00378** around 0.88140 — **wider than the SESOI (0.003)**, which is exactly why a fold-0 gap alone cannot conclude and the pooled CV of Table 3 is the endpoint.

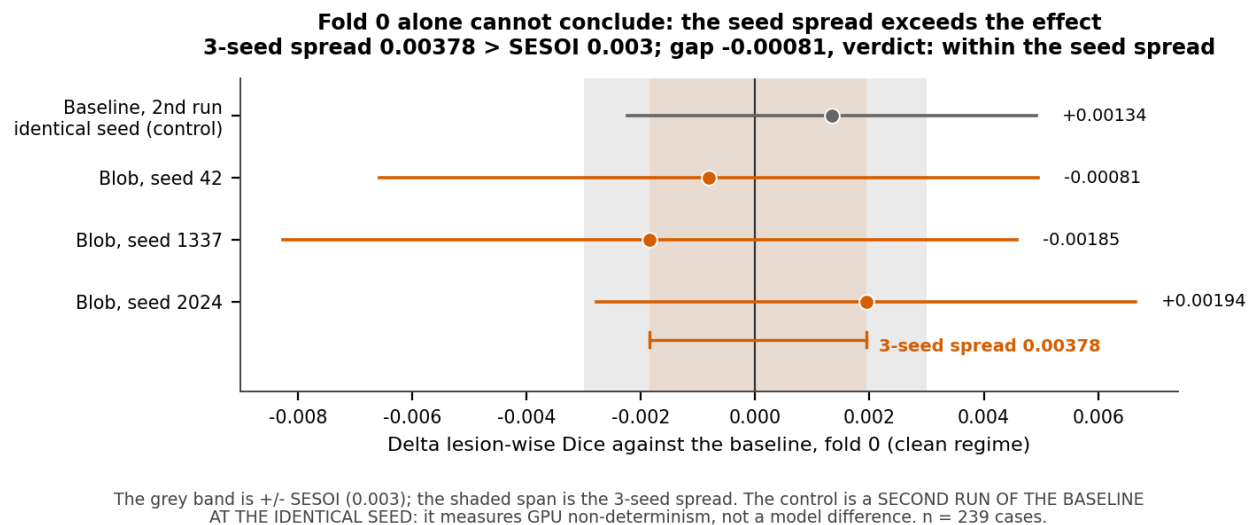


Figure 5: Fold 0 as an instrument rather than as a result: the three blob seeds, the same-seed baseline re-run that measures GPU non-determinism, and the seed spread drawn as a range against the SESOI. The spread exceeds the effect of interest, so no fold-0 gap can conclude.

4.8 Where the arm does pay: as a corroborator in an ordered veto

Table 9 ranks the twelve ordered two-expert vetoes on the primary endpoint, Holm over the twelve. `cc_X_Y` keeps the connected components of X that Y corroborates. The best is `cc_B_G` — the baseline’s components, vetoed by the blob arm — at +0.001137 (95% CI [0.000098, 0.002175], raw p = 0.0173, Holm p = 0.156, MDE 0.00148), rank 1 of 12. The runner-up is `cc_B_K` at +0.000703 (Holm p = 0.222).

Three things bound this number, and the paper states all three every time it states the number. It is **below the SESOI** (0.003), like all 12 vetoes of the table. It does **not survive Holm** (p = 0.156). 5 of the twelve are nominally significant before correction, and the 3 that survive it — `cc_G_K`, `cc_G_B`, `cc_G_D` — are **all negative**, all three of them operators that keep the blob arm’s own components; not one positive delta in this table survives Holm. And it is **asymmetric**: reversing the order gives `cc_G_B` at -0.002306, rank 11 (Holm p = 3.9e-14). Only 4 of the twelve vetoes are positive at all, and the four that are all keep the *baseline*’s components.

That asymmetry is the interpretable part. A veto that keeps X’s components and asks Y to corroborate them can only remove X’s false positives; it can never add Y’s discoveries. When `cc_B_G` beats `cc_G_B` by the margin it does, the measurement is saying that the blob arm’s components are worth *checking* the baseline’s against and not worth *keeping*. The component counts agree: after the veto the detection rate is the baseline’s (0.9226 against 0.9226) and the false positives fall from 57 to 47. The arm is a corroborator. Its value is as a second opinion that costs a training run, which is a poor trade on its own — and a cheap one inside a programme that trains one arm per auxiliary term anyway and scores them all on the same population.

`cc_B_G` (baseline components vetoed by the blob arm) is **rank 1 of 12** at +0.001137, CI [0.000098, 0.002175], Holm p = 0.1556 — **not significant and below the SESOI**, and it must be reported as such. The veto is **asymmetric**: reversing the order (`cc_G_B`) drops it to rank 11 with -0.002306. What survives is the baseline’s component, not the blob’s: the blob arm is useful as a **corroborator**, not as a **segmenter**.

Table 9 — Where the blob arm does pay: as a veto member

Twelve ordered 2-expert vetoes, same population and regime, scored on the mean of the three regions. `cc_X_Y` keeps the connected components of X that Y corroborates. Ranks are by delta against the baseline. Holm family: the 12 vetoes.

Rank	Veto (2 experts)	Delta Dice (mean)	95% CI	p	p (Holm)	MDE	Cases won	Below SESOI
1	<code>cc_B_G</code> — B vetoed by G	+0.001137	[0.000098, 0.002175]	0.0173	0.1556	0.00148	7	yes
2	<code>cc_B_K</code> — B vetoed by K	+0.000703	[-0.000025, 0.001431]	0.0277	0.2217	0.00104	6	yes
3	<code>cc_B_D</code> — B vetoed by D	+0.000673	[-0.000056, 0.001403]	0.063	0.4409	0.00104	5	yes
4	<code>cc_D_G</code> — D vetoed by G	+0.000152	[-0.002063, 0.002368]	0.2653	0.7959	0.00317	614	yes
5	<code>cc_D_B</code> — D vetoed by B	-0.000299	[-0.002365, 0.001766]	0.3249	0.7959	0.00295	612	yes
6	<code>cc_D_K</code> — D vetoed by K	-0.000361	[-0.002556, 0.001835]	0.3048	0.7959	0.00314	614	yes
7	<code>cc_K_B</code> — K vetoed by B	-0.000613	[-0.002677, 0.001451]	0.1037	0.622	0.00295	644	yes
8	<code>cc_K_D</code> — K vetoed by D	-0.000788	[-0.002987, 0.001411]	0.1457	0.6512	0.00314	642	yes
9	<code>cc_K_G</code> — K vetoed by G	-0.000832	[-0.003240, 0.001576]	0.1302	0.6512	0.00344	644	yes
10	<code>cc_G_K</code> — G vetoed by K	-0.002294	[-0.004768, 0.000179]	3.7e-14	3.7e-13	0.00354	481	yes
11	<code>cc_G_B</code> — G vetoed by B	-0.002306	[-0.004765, 0.000153]	3.2e-15	3.9e-14	0.00352	478	yes
12	<code>cc_G_D</code> — G vetoed by D	-0.002406	[-0.004851, 0.000038]	1.6e-14	1.7e-13	0.00349	480	yes

4.9 Reproducibility and controls

The consolidation that produces every number above runs 36/36 internal controls, all green, listed in Table 10 with the value each one verified. Two of them are back-validations against an earlier, independent scoring of the *same* checkpoints by the cross-validation harness, asserted at 1e-9: the pooled delta of this arm, recorded there as -0.003756015478, is recomputed here with a gap of 8.67e-19, and the `cc_B_G` veto delta, recorded as 0.001136680289, with a gap of 2.17e-19. This paper therefore cannot tell a different story from that earlier reading of the same artefacts — if it tried, a control would fail loudly rather than silently.

5. Discussion

The sign of an instance-wise loss is a property of the dataset. Kofler et al.’s mechanism — false positives enter the denominator of every instance, so their penalty scales with instance multiplicity — predicts a precision gain where instances are numerous and small, and it is what their multiple-sclerosis cohort shows. Adult glioma is the other case: a handful of large components with small satellites and annotation artefacts around them. There, making each instance count equally does not suppress spurious components; it makes the model *try* for every small structure the ground truth happens to contain, and the price is a 43 % rise in unmatched components. Our counts (1418 to 2029 in the raw regime) and the independent nnU-Net benchmark of Bouteille et al. on BraTS 2020 — instance-level precision 0.7159 against 0.7708, instance-wise F1 0.4789 against 0.4933, volumetric Dice flat at 0.9168 against 0.9171 (Bouteille et al. 2026) — agree on the direction from different data, a different backbone configuration and a different epoch budget. Kundu et al.’s positive result on BraTS-METS 2025 (foreground Dice 0.615 against a 0.59 baseline, resection cavity 0.150 to 0.221) sits on the other side of the same divide: many small instances, precision-positive (Kundu et al. 2026). Three datasets, two signs, one explanatory variable.

The evaluation protocol is not a neutral wrapper. 93.2 % of the Dice gap and 96.2 % of the HD95 gap exist only in the raw regime; the official component filter removes them. A method paper that reports raw predictions is reporting a quantity no challenge scores, and one that reports only cleaned predictions cannot show that its method changed the component-size distribution at all. This is the concrete, measured version of a complaint the metrics literature has made in general terms (Maier-Hein et al. 2024 ; Reinke et al. 2024): the post-processing is part of the method, and the honest report carries both regimes plus the share one absorbs of the other.

Aligning the loss with the metric is not the same as optimising it. The blob term mimics the ranking metric’s instance-wise structure more closely than any other term we have trained, and it is the only one of the three measurably worse under that metric. The reason is visible in the counts: the metric scores *matched* components, and it scores a false positive as a whole missed lesion. A loss that equalises the weight of every instance therefore equalises the weight of every *spurious* instance too, and on a dataset where the ground truth has few components, the extra ones the model produces are not

Table 10 — Controls executed by the consolidation itself

36/36 green. Two of them are back-validations against an earlier, independent scoring of the *same* checkpoints, at 1e-9: this paper cannot tell a different story from that reading of the same artefacts. The five identities that are re-asserted on every population and regime are grouped here into one row each, carrying their scope and their verdict; the value each individual run verified is in `analysis/blobloss_stats.json` (`controles`), which is what the deposit ships.

Control run by the consolidation	Scope	Verdict	Detail
Fold 0: 7200 measures (5 arms × 2 regimes × 3 regions × 240 cases)	—	OK	7200
Pooled CV: 200928 rows = 28 arms × 2 regimes × 3 regions × 1196 cases	—	OK	200928 rows, expected 200928
Pooled CV: 239 predictions at folds 1-4, 240 at fold 0 (documented gap)	—	OK	{0: 240, 1: 239, 2: 239, 3: 239, 4: 239}
Pooled CV: 1196 pooled cases, disjoint folds	—	OK	{'tous': 1196, 'sans_degenere': 1195}
7 runs at 300 completed epochs (read from the training logs)	—	OK	all 7 = 300
beta = 0.5, lambda_dist = 0.0 on all 7 runs (from debug.json)	—	OK	betas={0.5} lambdas={0.0}
Campaign closed 7/7, 0 failure	—	OK	{'runs': 7, 'completed': 7, 'failures': 0}
Back-validation: pooled delta of this arm against an earlier independent scoring of the same checkpoints (1e-09)	—	OK	recorded -0.003756015478 vs recomputed -0.003756015478 (gap 8.67e-19)
Back-validation: cc_B_G veto delta against that same earlier scoring (1e-09)	—	OK	recorded +0.001136680289 vs recomputed +0.001136680289 (gap 2.17e-19)
5 folds at 300 epochs, as recorded by the campaign	—	OK	{'0': 300, '1': 300, '2': 300, '3': 300, '4': 300}
Fold 0, clean: verdict taken verbatim from summary.json	—	OK	indistinguishable (within the 3-seed spread)
Fold 0, raw: verdict taken verbatim from summary.json	—	OK	outside the control's noise
3-seed spread exceeds the SESOI in clean (a fold-0 gap alone concludes nothing)	—	OK	spread 3 seeds = 0.0038, SESOI = 0.003
SESOI 0.003 recorded in the JSON	—	OK	0.003
Holm family of the primary criterion = the 3 regions (pre-reg. §3)	4 runs: 2 populations × 2 regimes	4/4 OK	3 tests
Paired num_fn == missed GT lesions of the pooled count (arm G)	4 runs: 2 populations × 2 regimes	4/4 OK	paired 306 vs pooled count 306 · paired 129 vs pooled count 129
Same identity for the reference B	4 runs: 2 populations × 2 regimes	4/4 OK	paired 310 vs pooled count 310 · paired 190 vs pooled count 190
Paired counts cover the population (lower + higher + unchanged == n)	4 runs: 2 populations × 2 regimes	4/4 OK	15 + 10 + 1171 vs n = 1196 · 37 + 2 + 1157 vs n = 1196 · ... (4 runs, all listed in the JSON)
Paired num_fp == false positives of the pooled count (arm G)	4 runs: 2 populations × 2 regimes	4/4 OK	paired 73 vs pooled count 73 · paired 2030 vs pooled count 2030 · ... (4 runs, all listed in the JSON)
Primary criterion (paired count of lesions at Dice = 0) computed in both regimes	2 runs: both populations	2/2 OK	['clean', 'raw']

The md5 of every artefact the consolidation reads is recorded inside `analysis/blobloss_stats.json` and listed file by file in the deposit's MANIFEST.sha256, so that any reader can check that the numbers above were computed from the artefacts they claim.

It is not reproduced as a table here: a fingerprint list is a property of the deposit, not an argument of the paper.

absorbed by averaging. Metric alignment transfers the metric's asymmetries into the gradient — it does not inherit its ranking.

What would have made this arm useful. Two things this paper cannot test. First, the weight: $\beta = 1.0$ was announced and never trained, and the raw-regime false-positive count is exactly the quantity a smaller β would trade against recall, so the deficit may be a β deficit rather than a mechanism deficit. Second, the decoding: a size-conditional filter — one that keeps small components only where the blob arm and the baseline agree — is the natural way to spend the asymmetry

measured in §4.8, and `cc_B_G` is its deterministic first approximation at +0.001137, under the SESOI. Neither is a claim; both are the next experiments this result implies, and both are declared as untested here.

Why publish it. The auxiliary-loss family measured on this backbone so far contains two nulls (Papers 1 and 2) and one pre-registered directional refutation (this one). Only the last carries a prediction that was frozen before the data and contradicted by it, which makes it the only one that can fail in an informative way (Varoquaux et al. 2022 ; Berger 1982). The practical instruction it leaves is narrow and usable: on adult glioma under the official BraTS-2023 protocol, do not add a blob term to Dice + CE to recover small lesions — it recovers them and loses more than it recovers; if the arm is already trained, its components are worth using to *veto* someone else’s.

6. Limitations

The loss weight is unmeasured. One β was trained (0.5). The $\beta = 1.0$ arm listed in the pre-registration was never launched, so every statement about the magnitude of the deficit is conditional on that single weight. This is the paper’s largest limitation and it is a choice of ours, not an accident: the arm was deprioritised while the campaign was running and the pre-registration was not amended.

One backbone, one dataset, one decoding family. MedNeXt-B under nnU-Net v2 on BraTS 2023 adult glioma, with the official per-region component thresholds. Nothing here generalises to a different architecture, to a different pathology — the metastases and multiple-sclerosis literature points the other way — or to a different cleaning policy.

No lesion-size stratification in the official regime. The claim that the arm recovers *small* lesions is supported by the counts (190 to 129 missed in the raw regime) but not by a per-size breakdown, because the only lesion-size distribution in the repository comes from the removed post-hoc harness and is measured on a proxy population (fit 180 / reserved 60). It is excluded rather than mixed into an official-scale table, which means the mechanism is inferred from totals.

No STAPLE, no learned combination. Probabilistic fusion (Warfield et al. 2004) has no verified implementation in this repository, and gates learned on cached expert probabilities were removed from it on 2026-08-07 after they produced rankings the official metric did not confirm. The twelve operators examined here are deterministic and ordered, and their family is declared; nothing learned is scored.

Cross-validation folds are not independent samples. Pooling 1195 cases over five folds treats them as one population for the paired test, which understates the variance coming from the fold split (Nadeau et al. 2003). The per-fold decomposition (Table 7) is reported for that reason: it shows a constant sign and a spread of 0.00346, and it is the strongest statement the design supports.

The paired mean is blunt. MDE 0.00390 against a delta of -0.00376: the rejection comes from the signed ranks, and a reader who wants magnitude rather than frequency should read the interval, not the p-value. Conversely, 1 row of the pre-registered criterion is significant after Holm with an interval containing zero — reported as such rather than dropped.

A stale artefact nearly became a claim. The `debug.json` of fold 0 reports `current_epoch = 15` where the training log shows 300 completed epochs; the epoch count in Table 1 is read from the log only, and the consolidation asserts it against the per-fold epoch counts recorded by the training campaign’s own artefacts. Any number in this paper that could come from two sources comes from the one that was checked.

7. Conclusion

A blob loss pre-registered in blind, trained for 300 epochs on five folds and three seeds of BraTS-2023 adult glioma, and scored only by the official lesion-wise capsule, does not do what its source literature predicted. It was expected to buy precision; it buys recall — 190 missed ground-truth lesions down to 129 in the raw regime, significant in 3/3 regions after Holm — and pays for it with 2029 lesion false positives against 1418, significant in 3/3 regions. Its lesion-wise Dice was expected to be neutral; it is -0.00376 (CI [-0.00648, -0.00103], Holm $p = 8.0e-14$), 1.25 times the pre-specified effect of interest, negative on 5 folds of 5, and the only one of the programme’s three auxiliary terms whose interval excludes zero. The official post-processing absorbs 93.2 % of that gap and 96.2 % of the HD95 gap, which is why the raw and clean regimes are reported together everywhere. The arm’s single positive contribution is as a corroborator — the ordered veto `cc_B_G` at +0.001137, rank 1 of 12, under the SESOI and not significant after Holm, and asymmetric because what survives a veto is the component of the arm that is kept. The generalisable statement is about the loss

family rather than about this dataset: the sign of an instance-wise loss follows the instance-size distribution of the data, and aligning an objective with a ranking metric transfers that metric’s asymmetries into the gradient instead of inheriting its ranking.

Author contributions

Guillaume Cassez — lead author: study design, pre-registration, model training, official evaluations, statistical analyses, consolidation and control pipeline, manuscript writing. **Stanislas Larnier** — formulation of the research questions, methodological guidance and careful review of successive drafts. Same roles as in Papers 1, 2 and 3.

Data, code and reproducibility

Trained on BraTS 2023 adult glioma (public challenge data). Every number in this paper is produced by a deterministic chain that reads only committed artefacts: `scripts/paper5_collector_chiffres.py` → `analysis/blobloss_stats.json` (36/36 controls, 16 sources with md5) → `scripts/paper5_tables.py` → the four `table_paper5` files in `tables/` → `scripts/paper5_figures.py` → `figures/` with `EM-PREINTES.json` → `scripts/paper5_manuscrit.py` → this file and its French mirror, with every marker traced in `tables/manuscrit_trace.json`. `scripts/paper5_coherence.py` re-reads the whole chain fail-closed. Metrics are computed by the official BraTS-2023 capsule (`external/BraTS-2023-Metrics`) and never by a local reimplementation. No stage of this pipeline requires a GPU.

References

- Abe, Buchanan, Pleiss, Zemel, Cunningham (2022). *Deep Ensembles Work, But Are They Necessary?*. Advances in Neural Information Processing Systems (NeurIPS).
- Baid, Ghodasara, Mohan, Bilello, Calabrese, Colak et al. (2021). *The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification*. arXiv preprint arXiv:2107.02314.
- Bakas, Akbari, Sotiras, Bilello, Rozycki, Kirby et al. (2017). *Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features*. Scientific Data. doi:10.1038/sdata.2017.117.
- Berger (1982). *Multiparameter Hypothesis Testing and Acceptance Sampling*. Technometrics. doi:10.1080/00401706.1982.10487790.
- Bouteille, Jaus, Kleesiek, Stiefelhagen, Heine (2026). *Learning to Look Closer: A New Instance-Wise Loss for Small Cerebral Lesion Segmentation*. IEEE International Symposium on Biomedical Imaging (ISBI).
- Bouthillier, Delaunay, Bronzi, Trofimov, Nichyporuk, Szeto et al. (2021). *Accounting for Variance in Machine Learning Benchmarks*. Proceedings of Machine Learning and Systems (MLSys).
- Christodoulou, Reinke, André, Godau, Kalinowski, Maier-Hein (2025). *False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims*. arXiv preprint arXiv:2505.04720. doi:10.48550/arXiv.2505.04720.
- Dangi, Linte, Yaniv (2019). *A distance map regularized CNN for cardiac cine MR image segmentation*. Medical Physics. doi:10.1002/mp.13853.
- Ding, Han, Niethammer (2019). *VoteNet: A Deep Learning Label Fusion Method for Multi-atlas Segmentation*. Medical Image Computing and Computer Assisted Intervention (MICCAI). doi:10.1007/978-3-030-32248-9_23.
- Holm (1979). *A Simple Sequentially Rejective Multiple Test Procedure*. Scandinavian Journal of Statistics.
- Isensee, Jaeger, Kohl, Petersen, Maier-Hein (2021). *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation*. Nature Methods. doi:10.1038/s41592-020-01008-z.
- Isensee, Jaeger, Full, Vollmuth, Maier-Hein (2020). *nnU-Net for Brain Tumor Segmentation*. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2020).
- Kervadec, Bouchtiba, Desrosiers, Granger, Dolz, Ben Ayed (2019). *Boundary loss for highly unbalanced segmentation*. Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning (MIDL).
- Kervadec, Bouchtiba, Desrosiers, Granger, Dolz, Ben Ayed (2021). *Boundary loss for highly unbalanced segmentation*. Medical Image Analysis. doi:10.1016/j.media.2020.101851.
- Kofler, Shit, Ezhov, Fidon, Horvath, Al-Maskari et al. (2023). *blob loss: Instance Imbalance Aware Loss Functions for Semantic Segmentation*. Information Processing in Medical Imaging (IPMI). doi:10.1007/978-3-031-34048-2_58.

- Kofler, Möller, Buchner, de la Rosa, Ezhov, Rosier et al. (2023). *Panoptica – instance-wise evaluation of 3D semantic and instance segmentation maps*. arXiv preprint arXiv:2312.02608. doi:10.48550/arXiv.2312.02608.
- Kundu, Kofler, Ivory, Moller, Shapey, Vercauteren (2026). *Instance Awareness of Multi-class Semantic Segmentation Loss Functions*. CV4CLINIC workshop at CVPR.
- Ma, Wei, Zhang, Wang, Lv, Zhu et al. (2020). *How Distance Transform Maps Boost Segmentation CNNs: An Empirical Study*. Proceedings of the Third Conference on Medical Imaging with Deep Learning (MIDL).
- Maier-Hein, Eisenmann, Reinke, Onogur, Stankovic, Scholz et al. (2018). *Why rankings of biomedical image analysis competitions should be interpreted with care*. Nature Communications. doi:10.1038/s41467-018-07619-7.
- Maier-Hein, Reinke, Godau, Tizabi, Buettner, Christodoulou et al. (2024). *Metrics reloaded: recommendations for image analysis validation*. Nature Methods. doi:10.1038/s41592-023-02151-z.
- Menze, Jakab, Bauer, Kalpathy-Cramer, Farahani, Kirby et al. (2015). *The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)*. IEEE Transactions on Medical Imaging. doi:10.1109/TMI.2014.2377694.
- Milletari, Navab, Ahmadi (2016). *V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation*. 2016 Fourth International Conference on 3D Vision (3DV). doi:10.1109/3DV.2016.79.
- Moawad, Janas, Baid, Ramakrishnan, Saluja, Ashraf et al. (2023). *The Brain Tumor Segmentation (BraTS-METS) Challenge 2023: Brain Metastasis Segmentation on Pre-treatment MRI*. arXiv preprint arXiv:2306.00838.
- Nadeau, Bengio (2003). *Inference for the Generalization Error*. Machine Learning. doi:10.1023/A:1024068626366.
- Nikolov, Blackwell, Zverovitch, Mendes, Livne, De Fauw et al. (2021). *Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study*. Journal of Medical Internet Research. doi:10.2196/26151.
- Picard (2021). *Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision*. arXiv preprint arXiv:2109.08203. doi:10.48550/arXiv.2109.08203.
- Reinke, Tizabi, Baumgartner, Eisenmann, Heckmann-Nötzel, others (2024). *Understanding metric-related pitfalls in image analysis validation*. Nature Methods. doi:10.1038/s41592-023-02150-0.
- Roy, Koehler, Ulrich, Baumgartner, Petersen, Isensee et al. (2023). *MedNeXt: Transformer-Driven Scaling of ConvNets for Medical Image Segmentation*. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. doi:10.1007/978-3-031-43901-8_39.
- Saluja, others (2023). *BraTS-2023-Metrics: Official BraTS 2023 Segmentation Performance Metrics*. .
- Sudre, Li, Vercauteren, Ourselin, Jorge Cardoso (2017). *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA/ML-CDS 2017). doi:10.1007/978-3-319-67558-9_28.
- Theisen, Kim, Yang, Hodgkinson, Mahoney (2023). *When are ensembles really effective?*. arXiv preprint arXiv:2305.12313. doi:10.48550/arXiv.2305.12313.
- Varoquaux, Cheplygina (2022). *Machine learning for medical imaging: methodological failures and recommendations for the future*. npj Digital Medicine. doi:10.1038/s41746-022-00592-y.
- Warfield, Zou, Wells (2004). *Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation*. IEEE Transactions on Medical Imaging. doi:10.1109/TMI.2004.828354.
- Wilcoxon (1945). *Individual Comparisons by Ranking Methods*. Biometrics Bulletin. doi:10.2307/3001968.
- Wolpert (1992). *Stacked generalization*. Neural Networks. doi:10.1016/S0893-6080(05)80023-1.
- Xue, Tang, Qiao, Gong, Yin, Qian et al. (2020). *Shape-Aware Organ Segmentation by Predicting Signed Distance Maps*. Proceedings of the AAAI Conference on Artificial Intelligence. doi:10.1609/aaai.v34i07.6946.
- Åkesson, Töger, Heiberg (2024). *Random effects during training: Implications for deep learning-based medical image segmentation*. Computers in Biology and Medicine. doi:10.1016/j.compbiomed.2024.108944.