

Sommaire

Elle détecte plus, elle score moins : une blob loss pré-enregistrée réfute sa propre prédiction de précision et ne paie que comme veto de consensus sous les métriques officielles BraTS-2023

Résumé	2
1. Introduction	2
2. Travaux connexes	4
3. Méthode	5
3.1 Ce qui était figé avant la mesure	5
3.2 Données, backbone et loss	5
3.3 Évaluation : une capsule, deux régimes de décodage	6
3.4 Statistique	6
3.5 Le plancher de bruit	7
4. Résultats	7
4.1 Le critère primaire pré-enregistré : H1 réfutée dans ses deux clauses	7
4.2 Endpoint primaire : un déficit constant en signe et réel en fréquence	9
4.3 Trois termes auxiliaires, un endpoint : seule la blob sort du bruit	10
4.4 Brut contre clean : le post-traitement est l'étage qui décide	11
4.5 Le mécanisme en comptes	12
4.6 Par fold : un signe constant sur cinq populations indépendantes	13
4.7 Le fold 0 seul : trois graines et un témoin de non-déterminisme	14
4.8 Là où le bras paie : comme corroborateur dans un veto ordonné	15
4.9 Reproductibilité et contrôles	16
5. Discussion	16
6. Limites	18
7. Conclusion	19
Contributions des auteurs	19
Données, code et reproductibilité	19
Références	19

Elle détecte plus, elle score moins : une blob loss pré-enregistrée réfute sa propre prédiction de précision et ne paie que comme veto de consensus sous les métriques officielles BraTS-2023

Guillaume Cassez · Stanislas Larnier

Recherche indépendante

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

Stanislas Larnier — stanislaslarnier@gmail.com · HAL stanislaslarnier

Version 1 — premier jet complet, écrit depuis le corpus consolidé le 2026-09-30T14:07:45. Troisième terme du programme BRATS sur les loss auxiliaires à être mesuré sur la validation croisée complète en cinq folds, après les bras à distance signée et à boundary loss. Tout y est lu du corpus consolidé que la **Provenance**, juste ci-dessous, nomme fichier par fichier : 16 artefacts dont le md5 est inscrit (§4.9), sous 36/36 contrôles internes. Cette prose ne porte **aucun chiffre tapé à la main** : elle est assemblée depuis un gabarit où chaque valeur est un marqueur résolu depuis ce JSON, et l'assemblage est déterministe — le régénérer reproduit ce fichier octet pour octet (--verif).

Convention de ce miroir. La prose et les tableaux sont en français. Les tableaux ne sont pas traduits après coup : scripts/paper5_tables.py parcourt le même analysis/blobloss_stats.json deux fois, une par langue, et écrit table_paper5.md (anglais) et table_paper5_fr.md (français) — un seul générateur, donc une seule source par nombre, ce qui est précisément ce que ce dossier exige. Deux garde-fous fail-closed sont rejoués à chaque génération : aucun mot-outil anglais ne peut survivre dans le

fichier français, et le multiset des jetons numériques des deux langues doit être identique (traduire des mots ne déplace aucun chiffre). Les séparateurs décimaux de la prose sont français ; ceux des tableaux restent le point, comme dans la racine anglaise et dans les citations verbatim d’artefacts — un tableau est la trace lisible d’un JSON, pas une phrase.

Provenance. Les affirmations quantitatives viennent de `scripts/paper5_collecter_chiffres.py` (`analysis/blobloss_stats.json`) ; les tableaux de `scripts/paper5_tables.py`, qui écrit `table_paper5.md` (anglais), `table_paper5_fr.md` (miroir français), `.tex` et `.json` dans `tables/`, ce dernier portant le chemin JSON de chaque valeur écrite ; les figures de `scripts/paper5_figures.py` dans `figures/`, avec le manifeste `EMPREINTES.json`. Les métriques sont calculées par la capsule officielle BraTS-2023 (`external/BraTS-2023-Metrics`) (Saluja et al. 2023), jamais par une réimplémentation locale. L’ensemble complet des métriques officielles est rapporté ; le critère primaire pré-enregistré est séparé de l’endpoint primaire, et les deux des analyses exploratoires ; le résultat négatif est dit dans le titre, le résumé et la conclusion.

Avertissement de périmètre (porteur). Trois choses que ce papier n’est **pas**. Ce n’est pas une étude du poids de la loss : seul $\beta = 0,5$ a été entraîné, et le bras $\beta = 1,0$ annoncé au §4 du pré-enregistrement n’a jamais été lancé — la sensibilité à ce poids n’est donc **pas mesurée** (§6). Ce n’est pas une étude d’une combinaison apprise : les seuls opérateurs examinés sont les douze vetos déterministes ordonnés à deux experts. Rien n’est appris sur des probabilités d’experts mises en cache, et un mélangeur appris que ce bras pourrait alimenter comme expert parmi d’autres — un MoE patch-wise en est le candidat naturel — relève du travail à venir (§5), pas d’un résultat de ce papier. Et ce n’est pas une étude en métrique proxy : les comptages lésionnels du harnais post-hoc supprimé (`outputs/pruning/lesions_*,gt_lesion_sizes.json`, fit 180 / reserved 60) sont exclus de tous les tableaux et figures, parce que mélanger deux échelles d’évaluation est la manière dont un programme perd le compte de ses propres nombres.

Résumé

Les loss par instance sont présentées comme le remède à la sous-segmentation des petites lésions : elles scorent chaque lésion séparément, de sorte qu’une petite instance manquée coûte autant qu’une grande. Nous avons pré-enregistré un test de l’une d’elles — la blob loss (Kofler et al. 2023) — sur le gliome adulte BraTS-2023 **avant de lire le moindre chiffre lésion-wise**, en prédisant, d’après la cohorte sclérose en plaques de l’article source, qu’elle déplacerait une baseline MedNeXt-B / nnU-Net (Roy et al. 2023 ; Isensee et al. 2021) vers la **précision** : pas de baisse des lésions GT à Dice = 0, et moins de faux positifs lésionnels. Les deux clauses sont réfutées, en sens opposés. Sur 1195 cas held-out poolés sur cinq folds disjoints et scorés par la capsule officielle BraTS-2023, le bras blob manque 129 lésions GT contre 190 pour la baseline (significatif dans 3/3 régions après Holm) et produit 2029 faux positifs lésionnels contre 1418 (3/3 régions) : il **achète du rappel et le paie en précision**, l’inverse exact de la prédiction. Le post-traitement officiel absorbe ensuite 93,2 % de l’écart de Dice brut et 96,2 % de l’écart de HD95 brut, laissant -0,00376 de Dice lésion-wise (IC 95 % [-0,00648 ; -0,00103], p de Holm = 8,0e-14) — soit 1,25 fois le plus petit effet d’intérêt pré-spécifié (0,003), négatif sur 5 folds sur 5. Des trois termes auxiliaires que nous avons mesurés sur ce backbone, la blob est le seul dont l’intervalle de confiance exclut zéro, et il l’exclut par derrière : la régression de distance signée donne -0,00060 (p de Holm = 1) et la boundary loss -0,00116 (p de Holm = 0,9547), toutes deux indiscernables de la baseline. Le seul delta positif du bras est comme **corroborateur** : le veto ordonné `cc_B_G` se classe 1 sur 12 à +0,001137, sous l’effet d’intérêt et non significatif après Holm (p = 0,156), et l’inversion de l’ordre le fait tomber au rang 11 (-0,002306). Le résultat n’est pas que la loss échoue ; c’est qu’une prédiction directionnelle posée en aveugle a été contredite par la mesure, et que le signe d’une loss par instance est une propriété de la distribution de tailles d’instances du dataset, pas de la loss.

1. Introduction

Les loss de segmentation qui moyennent sur un volume traitent les lésions d’un patient comme une seule masse. Quand un examen porte quelques grosses composantes et beaucoup de petites, les petites ne contribuent presque rien au gradient, et un modèle peut toutes les perdre sans que son Dice bouge sensiblement (Kofler et al. 2023 ; Kofler et al. 2023). Les loss par instance suppriment cette moyenne : la blob loss score chaque composante GT contre la prédiction restreinte à cette composante, de sorte que chaque lésion pèse le même poids dans l’objectif, quel que soit son volume. La métrique qu’elle imite est exactement celle dont le protocole BraTS-2023 se sert pour classer les soumissions — Dice et HD95

lésion-wise, calculés par composante connexe (Saluja et al. 2023). La question n'est donc pas exotique : **si la loss est alignée sur la métrique de classement, le modèle bouge-t-il sur cette métrique, et dans quel sens ?**

Deux d'entre nous ont déjà posé la question deux fois sur ce backbone. Le papier 1 ajoutait une tête de régression de transformée de distance signée (Ma et al. 2020) et trouvait un régulariseur orienté **rappel** qui ne bat pas la baseline ; le papier 2 ajoutait une boundary loss (Kervadec et al. 2019 ; Kervadec et al. 2021) et trouvait son miroir en **précision**, sans battre la baseline non plus. Ces deux résultats définissent un axe — rappel contre précision — et laissent un troisième terme non testé. La blob loss est ce troisième point naturel, et le plus intéressant, parce qu'elle est le seul des trois dont la construction est **par instance** plutôt que **par frontière** : elle n'affine pas les contours, elle re-pondère les lésions.

Notre propre déficit mesuré est du côté rappel de cet axe, et c'est ce qui a motivé le chantier : à la date du pré-enregistrement, la baseline manquait encore entièrement un jeu de lésions GT, et ces lésions manquées étaient de plusieurs ordres de grandeur plus petites que les détectées. La prédiction que nous avons figée était pourtant **contraire** à cette intuition, parce que la source primaire pointe dans l'autre sens. Dans la cohorte sclérose en plaques de Kofler et al. — le contexte le plus proche de « nombreuses petites lésions » — la sensibilité par instance **baisse** (0,854 → 0,822) pendant que la précision monte (0,711 → 0,828) — chiffres consignés par notre pré-enregistrement après une lecture complète de la source, et non re-vérifiés en ligne depuis — et leur §6.9 en donne le mécanisme : le terme blob met les faux positifs au dénominateur de *chaque* instance, donc pénalise les composantes parasites proportionnellement au nombre d'instances du cas. Sur leurs autres datasets (LiTS, SHANEL, DeepMACT) la sensibilité par instance monte au contraire. Le signe n'est pas fixé par la loss.

Nous avons donc pré-enregistré. docs/blob_loss_preregistrement.md a été rédigé le 2026-08-06 à 12h10, pendant que le premier run était à l'époque 11 sur 300 et **avant que le moindre chiffre lésion-wise ait été lu**. Il fige la question, une hypothèse directionnelle H1 (un déplacement vers la précision : les lésions à Dice = 0 ne baissent pas, les faux positifs lésionnels baissent), une hypothèse nulle H2 (Dice lésion-wise neutre, $|\Delta| < 0,003$), le critère primaire (le compte apparié par patient des lésions GT à Dice = 0, Wilcoxon signé avec correction de Holm sur les trois régions) et les garde-fous (décodage officiel S1, comparaison appariée contre la baseline du même fold et de la même graine, au moins trois graines dès que l'écart observé passe sous 0,005). Un cadrage écrit après lecture n'aurait aucune valeur probante ; celui-ci est antérieur, et le §4.1 le reproduit tel quel.

Ce que dit la mesure est court à énoncer et désagréable à avoir prédit. **H1 est réfutée dans ses deux clauses, et en sens opposés** : en régime de décodage brut, le bras blob manque 129 lésions GT contre 190 — une amélioration, significative dans 3/3 régions après Holm — tout en produisant 2029 faux positifs lésionnels contre 1418, significatifs eux aussi dans 3/3 régions. **H2 est réfutée aussi, du mauvais côté** : le delta poolé de Dice lésion-wise vaut -0,00376 avec un intervalle de confiance qui exclut zéro. La loss fait ce que dit sa construction — elle récupère les petites instances — et le paie en composantes parasites, sur un dataset où les gliomes forment quelques grosses composantes entourées de petits satellites. Le nettoyage officiel de composantes retire ensuite l'essentiel des deux mouvements, et ce qui survit est un faible déficit.

Contributions :

1. **Un résultat négatif pré-enregistré, sous métrique officielle.** Nous rapportons le critère primaire exactement tel qu'il a été figé, dans les deux régimes de décodage, avec l'endpoint dont le programme se sert pour classer (Tableau 2, Fig. 1). La prédiction était directionnelle et posée en aveugle ; elle est contredite ; nous publions la contradiction plutôt que le déficit, parce que le déficit est petit et la contradiction non.
2. **La mesure de ce qui décide du verdict.** Le post-traitement officiel absorbe 93,2 % de l'écart de Dice brut et 96,2 % de l'écart de HD95 brut (Tableau 5, Fig. 2). Publier le régime brut aurait annoncé un effondrement (-0,0554 en Dice, +20,6579 mm) que le protocole officiel ne mesure pas ; publier le seul régime clean aurait caché le mécanisme. Les deux sont montrés, et l'endpoint est déclaré une fois.
3. **Le mécanisme énoncé en comptes, pas en ratio.** Sur 4007 lésions GT, le bras blob en détecte 3878 contre 3817 pour la baseline, et laisse 2029 composantes prédites non appariées contre 1418 — soit 43 % de hausse (Tableaux 6a, 6b). Après nettoyage officiel l'avantage de détection a disparu (306 manquées contre 310) tandis que l'écart de faux positifs se resserre sans se refermer (72 contre 57).
4. **Un positif borné.** Le seul delta du bras au-dessus de la baseline est comme membre de veto : cc_B_G se classe 1 sur 12 vetos ordonnés à deux experts à +0,001137 (p de Holm = 0,156), sous l'effet d'intérêt, et le veto est asymétrique — inversé, il tombe au rang 11 à -0,002306 (Tableau 9). Ce qui survit au veto est la composante de la baseline, pas celle du bras blob : un corroborateur, pas un segmenteur.

5. **Un plancher de bruit qui interdit la lecture facile.** Trois graines au fold 0 s'étendent sur 0,00378 en Dice lésion-wise clean — plus que l'effet d'intérêt et plus que l'écart du fold 0 lui-même (-0,00081) — et un second run de la baseline à graine **identique** bouge de +0,00134 (Tableau 8, Fig. 4). Aucune conclusion de ce papier ne repose sur un fold seul.

2. Travaux connexes

Loss par instance. La blob loss (Kofler et al. 2023) isole chaque composante GT en masquant la prédiction sur elle, en gardant les faux positifs en jeu, et moyenne la métrique de base sur les composantes ; le cadre Panoptica des mêmes auteurs (Kofler et al. 2023) en est le pendant côté évaluation. La construction est ce qui rend son comportement dépendant du dataset : comme le score de chaque instance est divisé par un terme qui inclut tous les faux positifs, la pénalité des composantes parasites croît avec la multiplicité des instances. Deux réplifications indépendantes cadrent notre résultat. Bouteille et al. (Bouteille et al. 2026) comparent DiceCE, BlobDiceCE et leur CC-DiceCE dans nnU-Net sur cinq cohortes dont BraTS 2020 (369 cas) ; sur BraTS le bras blob tient le Dice volumétrique (0,9168 contre 0,9171) mais perd en précision au niveau instance (0,7159 contre 0,7708) et en F1 lésion-wise (0,4789 contre 0,4933), et les auteurs relèvent que les gliomes forment peu de composantes, très grandes, avec de petits artefacts d'annotation autour — la structure même que montrent nos comptes. Kundu et al. (Kundu et al. 2026) étendent la blob loss au multi-classe par décomposition one-vs-rest et rapportent un résultat **positif** sur BraTS-METS 2025 (260 cas test) : Dice foreground 0,615 contre une baseline à $0,59 \pm 0,27$, Dice de la cavité de résection relevé de 0,150 à 0,221, et la meilleure qualité panoptique au seuil 0,5 ($0,40 \pm 0,24$ contre $0,38 \pm 0,25$). Lues ensemble, les trois études couvrent trois régimes de tailles d'instances — beaucoup de lésions minuscules (métastases, positif), quelques grosses composantes avec petits satellites (gliome, négatif en précision dans les deux mesures indépendantes), et des petites lésions multiples de SEP (positif en précision dans l'article source). Le signe suit le régime, pas la loss.

Termes auxiliaires sur un backbone fort. La régression de distance signée (Ma et al. 2020 ; Xue et al. 2020 ; Dangi et al. 2019) et la boundary loss (Kervadec et al. 2019 ; Kervadec et al. 2021) sont les deux termes que les papiers 1 et 2 de ce programme ont mesurés sur le même backbone MedNeXt-B / nnU-Net v2 ; ils sont re-scorés ici sur la population identique et la famille de Holm identique comme **comparateurs** (Tableau 4), et non ré-entraînés. Les loss par région et les ensembles pour BraTS sont les alternatives standards (Isensee et al. 2020 ; Abe et al. 2022 ; Theisen et al. 2023), et les opérateurs de consensus entre modèles entraînés indépendamment sont la famille sans entraînement à laquelle appartient le veto de ce papier (Warfield et al. 2004 ; Wolpert 1992 ; Ding et al. 2019).

Backbone, données et métriques. nnU-Net (Isensee et al. 2021) avec le bloc MedNeXt-B (Roy et al. 2023) ; Dice et Dice généralisé (Milletari et al. 2016 ; Sudre et al. 2017) ; gliome BraTS (Menze et al. 2015 ; Bakas et al. 2017 ; Baid et al. 2021) ; les métriques officielles lésion-wise BraTS-2023 et leur complément de distance de surface (Saluja et al. 2023 ; Moawad et al. 2023 ; Nikolov et al. 2021).

Pourquoi le protocole est énoncé avant le résultat. Classer sur un fold unique, sur une graine unique, ou sur une métrique qui n'est pas celle du challenge, est le mode de défaillance documenté de cette littérature (Maier-Hein et al. 2018 ; Maier-Hein et al. 2024 ; Reinke et al. 2024 ; Christodoulou et al. 2025). La variance de graine seule peut dépasser un effet publié (Picard 2021 ; Bouthillier et al. 2021 ; Åkesson et al. 2024), les folds d'une validation croisée ne sont pas des échantillons indépendants (Nadeau et al. 2003), et un test d'hypothèse nulle ne peut pas établir une absence d'effet sans borne d'équivalence (Berger 1982). Les résultats négatifs méritent d'être publiés précisément parce que la littérature les sous-rapporte (Varoquaux et al. 2022). Chacun de ces points est opérationnalisé ci-dessous : un plus petit effet d'intérêt pré-spécifié, une correction de Holm à l'intérieur de familles **déclarées**, trois graines plus un témoin à graine identique, et une validation croisée poolée comme endpoint plutôt qu'un fold quelconque.

3. Méthode

3.1 Ce qui était figé avant la mesure

Le pré-enregistrement (docs/blob_loss_preregistrement.md, 2026-08-06 12h10) fixe quatre choses que ce papier n’a plus le droit de choisir après coup :

- **H1 (directionnelle).** Le terme blob déplace le modèle vers la *précision* : le nombre de lésions GT à Dice = 0 **ne baisse pas**, et les faux positifs lésionnels baissent. La prédiction a été écrite délibérément contre notre propre intuition, sur la foi de la cohorte sclérose en plaques de l’article source.
- **H2 (nulle).** Le Dice lésion-wise global est neutre : $|\Delta| < \text{SESOI} = 0,003$.
- **Critère primaire.** Le compte apparié par patient des lésions GT à Dice = 0, Wilcoxon signé, Holm sur les trois régions (WT, TC, ET). Dans le scorer officiel BraTS-2023, une lésion GT à Dice = 0 est une lésion GT non appariée (num_fn) ; cette équivalence de définition n’est pas supposée, elle est **assertée** par la consolidation contre le comptage poolé, dans les deux populations et les deux régimes (Tableau 10).
- **Garde-fous.** Décodage officiel S1 ; comparaison appariée contre la baseline du même fold et de la même graine ; **au moins trois graines** dès que l’écart observé passe sous 0,005, parce que le bruit de graine mesuré sur une tête libre vaut $\text{sd} = 0,005$ à l’échelle voxel-wise, au-dessus de l’effet d’intérêt ; et aucune conclusion tirée d’un relevé court.

Deux de ces engagements ont été tenus, un ne l’a pas été entièrement. Trois graines ont été entraînées au fold 0, comme exigé. Le bras $\beta = 1,0$ listé au §4 du pré-enregistrement n’a jamais été lancé, donc la sensibilité au poids de toutes les affirmations ci-dessous est inconnue (§6). Le pré-enregistrement déclarait aussi que ce travail serait un **papier autonome** dans la famille des loss auxiliaires plutôt qu’une section ajoutée à une autre étude : c’est ce qu’est ce document.

3.2 Données, backbone et loss

BraTS 2023 gliome adulte, dataset nnU-Net de 1196 cas après les 55 exclusions documentées, découpé en cinq folds disjoints ; la validation held-out couvre 240 cas au fold 0 et 239 à chacun des folds 1 à 4. Un cas, BraTS-GLI-00017-001, porte une vérité terrain vide, pour laquelle le Dice lésion-wise est indéfini ; il est retiré de la population poolée, ce qui donne $n = 1195$, la population sur laquelle tous les autres bras du programme sont scorés. La population complète à $n = 1196$ est aussi calculée et ne change aucune conclusion.

Le backbone est MedNeXt-B sous nnU-Net v2 avec les plans nnUNetPlans_96GB_mednext, entraîné 300 époques sur une unique NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition. Le bras ne diffère de la référence que par un terme :

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}} + \beta \cdot \mathcal{L}_{\text{blob}}, \quad \beta = 0,5, \quad \lambda_{\text{dist}} = 0,0$$

avec $\beta = 0,5$ qui reproduit le rapport 2:1 global:blob de Kofler et al., et le terme de distance signée du papier 1 désactivé. Les deux paramètres sont relus du debug.json de **chaque** run plutôt que de la documentation, et le compte d’époques du journal d’entraînement — le debug.json du fold 0 porte un current_epoch périmé à 15 là où le journal montre 300 époques terminées, ce qui est exactement la raison pour laquelle les deux sources ne sont pas interchangeables. Sept runs ont été entraînés : cinq folds à la graine 42 et deux graines supplémentaires (1337, 2024) au fold 0. La campagne a été close le 2026-08-20T07:20:36Z avec 7 runs terminés sur 7 et 0 échec (Tableau 1).

Sur le nombre qui atteste la convergence. nnU-Net écrit un *Mean Validation Dice* en fin d’entraînement : 0,90622 en moyenne sur les cinq folds, avec une étendue de 0,00184 entre les trois graines du fold 0. Ce nombre est **voxel-wise**, calculé avec mirroring d’augmentation au temps du test et sans nettoyage de composantes. Ce n’est pas la métrique officielle BraTS, il n’est comparable à rien des Tableaux 2 à 9, et il n’apparaît dans ce papier que pour une seule chose : attester que chaque run a entraîné jusqu’au budget.

Tableau 1 — Ce qui a été entraîné (attestation de campagne)

Le Mean Validation Dice est le nombre **voxel-wise** que nnU-Net écrit en fin d’entraînement, avec TTA mirroring et sans nettoyage de composantes connexes. Ce n’est **pas** la métrique officielle BraTS et il n’est rapporté ici que pour attester que chaque run a convergé sur tout le budget. Tous les résultats officiels sont dans les Tableaux 2 à 9.

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

Campagne : **7 runs, 7 terminés, 0 échec**, close le 2026-08-20T07:20:36Z. Moyenne sur les 5 folds (graine 42) : **0.90622** ; étendue des 3 graines au fold 0 : **0.00184**.

3.3 Évaluation : une capsule, deux régimes de décodage

Toutes les métriques viennent de la capsule officielle BraTS-2023 ([external/BraTS-2023-Metrics](#)) à l’échelle **lésion-wise** : Dice et HD95 par région (WT, TC, ET), calculés sur les composantes connexes et non sur le volume. Deux régimes de décodage sont rapportés côte à côte partout, parce que leur différence est l’un des résultats de ce papier :

- **clean — le régime officiel.** Décodage argmax suivi de la suppression, par région, des composantes connexes de moins de 1000 voxels (WT), 250 (TC) et 500 (ET). C’est ce que le challenge score, et c’est le régime primaire déclaré de ce papier.
- **raw — aucun nettoyage.** La même prédiction argmax, sans filtrage de composantes. Il est rapporté parce que la blob loss est *conçue* pour modifier la distribution de tailles des composantes, et qu’un régime qui supprime les petites composantes avant de scorer cache exactement l’effet étudié.

La validation croisée poolée couvre 200928 lignes scorées (28 bras × 2 régimes × 3 régions × 1196 cas) ; l’évaluation du fold 0 ajoute 7200 mesures (5 bras × 2 régimes × 3 régions × 240 cas).

3.4 Statistique

Wilcoxon signé apparié patient par patient (Wilcoxon 1945), correction de Holm à l’intérieur de familles **déclarées** (Holm 1979) : la famille des experts {G, D, K} × {WT, TC, ET, moyenne} pour l’endpoint primaire ; les trois régions pour le critère primaire pré-enregistré, tel qu’il a été figé ; les douze vetos ; et 20 tests (5 folds × 4 régions), régime clean, métrique Dice lésion-wise pour la décomposition par fold. Les intervalles de confiance sont à 95 % sur la moyenne appariée ; l’effet minimal détectable est $MDE = 2,802 \times se$. Le plus petit effet d’intérêt est pré-spécifié à $SESOI = 0,003$, et le rapport $|\Delta|/SESOI$ est rapporté à côté de chaque delta de premier plan pour qu’un lecteur n’ait jamais à le reconstruire.

Deux instruments sont rapportés ensemble partout, parce que l’un ou l’autre cité seul induirait en erreur. La **moyenne** appariée est grossière sur ces données : sd pairé 0,0481 pour un delta de -0,00376, d’où une MDE nominale de 0,00390, supérieure à l’effet qu’elle est censée détecter. C’est le test des **rangs signés** qui rejette, et ce qu’il dit n’est pas « bien pire » mais « pire plus souvent que meilleur » : sur 1193 cas dont le score a bougé, 713 ont baissé. Le verdict publiable est la conjonction — *un déficit faible, constant en signe, réel en fréquence, d’une ampleur qui se situe autour du seuil d’intérêt clinique* — et aucun nombre de ce papier ne le porte seul.

Une ligne du critère pré-enregistré est significative après Holm alors que son intervalle sur la différence de moyennes contient encore zéro (1 sur 12 lignes, les deux régimes confondus). Ce n’est pas une contradiction à lisser : l’intervalle décrit la moyenne **appariée**, la p-value les **rangs signés**, et sur des comptes nuls pour presque tous les cas l’effet vit dans la fréquence à laquelle un compte bouge, pas dans l’ampleur dont il bouge en moyenne. Sur l’endpoint primaire, 2 des 4 lignes rapportées sont significatives avec un intervalle qui chevauche zéro (TC et ET) ; les deux sont montrées telles quelles dans le Tableau 3.

3.5 Le plancher de bruit

Un écart mesuré sur un fold ne s'interprète pas sans savoir ce que la même commande d'entraînement produit deux fois. Deux témoins sont entraînés pour cela :

- **trois graines** du bras blob au fold 0 (42, 1337, 2024), la réplication que le pré-enregistrement exigeait ;
- **un second run de la baseline à graine identique**, qui mesure le non-déterminisme GPU et rien d'autre — pas une différence de modèle.

Ils sont rapportés au Tableau 8 et en Fig. 4, avec l'écart du fold 0 entre les deux. Les verdicts écrits par l'évaluation sont ceux repris ici verbatim : indiscernable (dans l'étendue des 3 graines) en régime clean, hors du bruit du témoin en régime brut.

4. Résultats

4.1 Le critère primaire pré-enregistré : H1 réfutée dans ses deux clauses

Le critère avait été figé comme le compte apparié par patient des lésions GT à Dice = 0, avec Holm sur les trois régions, et H1 prédisait qu'il **ne baisserait pas** pendant que les faux positifs lésionnels baisseraient. Le Tableau 2 et la Fig. 1 le rapportent dans les deux régimes de décodage.

En régime **brut**, les deux clauses échouent, et elles échouent en sens opposés. Les lésions GT manquées passent de 190 à 129 — significatif dans 3/3 régions après Holm (WT Δ -0,02008, $p = 7,5e-06$; TC Δ -0,01590, $p = 6,8e-05$; ET Δ -0,01506, $p = 2,4e-04$), avec 37 cas améliorés contre 2 dégradés et 1156 inchangés. Les faux positifs lésionnels montent de 1418 à 2029, là aussi significatifs dans 3/3 régions (WT +0,27280, $p = 1,0e-22$; TC +0,17824, $p = 6,7e-21$; ET +0,06025, $p = 1,9e-06$), avec 408 cas dégradés contre 123 améliorés. La loss achète du rappel et le paie en précision — l'inverse de la prédiction, qui venait d'une cohorte SEP où la multiplicité des instances est élevée et où le dénominateur de faux positifs par instance fait le travail opposé.

En régime **clean**, le filtrage officiel de composantes absorbe l'essentiel du mouvement. Les lésions manquées passent de 310 à 306, significatif dans 1/3 régions (WT -0,00669, $p = 0,0342$; TC +0,00000, $p = 1$; ET +0,00335, $p = 0,3146$), avec 15 cas améliorés contre 10 dégradés et 1170 inchangés. Les faux positifs passent de 57 à 72, significatifs dans 0/3 régions. Ce que le régime brut montrait comme un grand déplacement bidirectionnel survit au protocole comme un petit — ce qui est le résultat du §4.4 en une ligne.

- **Régime brut.** Les lésions manquées tombent de **190** (baseline) à **129** (blob), significatif dans 3 régions sur 3 après Holm ($p = 7,5e-06$, $6,8e-05$, $2,4e-04$) ; 37 cas s'améliorent contre 2 qui se dégradent. Les faux positifs montent de **1418** à **2029**, significatif dans 3 régions sur 3. La loss fait l'**inverse** de la prédiction pré-enregistrée : elle achète du rappel et le paie en précision.
- **Régime clean.** Le nettoyage officiel en efface l'essentiel : lésions manquées **310** → **306**, encore significatif dans 1 région sur 3 (WT $p = 0,0342$) ; 15 cas s'améliorent contre 10 qui se dégradent, 1170 inchangés. Faux positifs **57** → **72**, 0 régions sur 3 significatives après Holm.
- **H2 telle que pré-enregistrée** prédisait un Dice lésion-wise *neutre* ($|\Delta| < \text{SESOI} = 0,003$). Mesuré (Tableau 3) : $\Delta = -0,00376$ avec un IC qui exclut zéro — **H2 est réfutée elle aussi**, du côté défavorable.
- **Deux instruments, déclarés ouvertement.** 1 ligne de ce tableau est significative après Holm alors que son IC de différence moyenne contient encore zéro (brut/faux_positifs/ET). Ce n'est pas une contradiction : l'IC décrit la **moyenne** appariée, la p-value les **rangs signés**. Sur des comptages nuls pour presque tous les cas, l'effet vit dans la fréquence à laquelle le comptage bouge, pas dans l'ampleur moyenne du mouvement. Les deux sont rapportés pour qu'aucun ne puisse être cité seul.

Conséquence pour le papier. Ce n'est pas un résultat nul déguisé en tel : une prédiction directionnelle faite en aveugle a été contredite par la mesure, et la contradiction est elle-même le résultat — aligner la loss sur la métrique lésion-wise fait bien bouger le modèle sur les comptages de lésions, simplement pas dans le sens que suggérait le dataset source (la cohorte SEP de Kofler), et le post-traitement officiel absorbe ensuite l'essentiel du mouvement.

H2 est réfutée elle aussi, du mauvais côté : le delta poolé de Dice lésion-wise vaut -0,00376 avec un intervalle de confiance excluant zéro (§4.2), là où H2 prédisait $|\Delta| < 0,003$. Aucune des deux hypothèses ne survit, et les deux échecs

Tableau 2 — Le critère primaire pré-enregistré, à l'épreuve

Source du pré-enregistrement : docs/blob_loss_preregistrement.md §3 (écrit le 2026-08-06, **avant toute lecture d'un nombre lésion-wise**). Critère tel qu'il y est figé, cité textuellement : *num_fn — lésions GT jamais apparées (Dice = 0), apparié par patient* — c'est le champ num_fn, compté par patient, donc le nombre de lésions

GT à Dice = 0 jamais apparées ; famille de Holm : 3 tests (les 3 régions WT/TC/ET), conformément au pré-enregistrement, soit 3 tests sur WT/TC/ET. L'équivalence avec les comptages poolés est attestée par le Tableau 10.

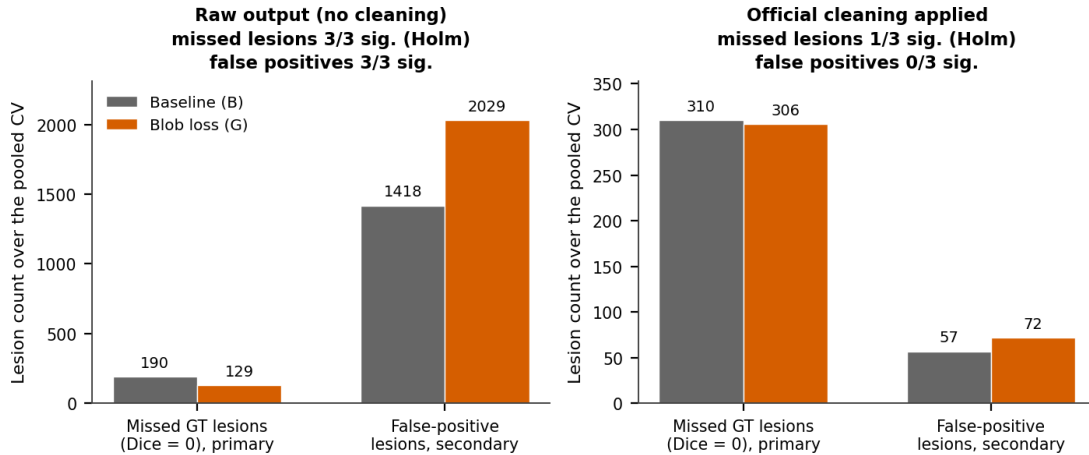
H1 telle que pré-enregistrée prédisait un déplacement vers la *précision* : le nombre de lésions à Dice = 0 ne devait **pas** baisser, et les faux positifs lésionnaires devaient baisser. **Mesuré : les deux clauses sont réfutées, dans des directions opposées.**

Lecture par groupe : le régime, le critère et les totaux poolés ne sont écrits qu'une fois par groupe, sur sa première et sa dernière ligne.

Régime	Critère pré-enregistré	Région	Delta par cas (G - B)	IC 95 %	p (Holm)	Cas où G gagne	Cas qui bougent	Total G	Total B
brut	PRIMAIRE — lésions GT manquées (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		
		poolé	—	—	—	37 en baisse / 2 en hausse / 1156 inchangés	1195	129	190
brut	Secondaire — lésions faux positifs	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		
		poolé	—	—	—	123 en baisse / 408 en hausse / 664 inchangés	1195	2029	1418
clean	PRIMAIRE — lésions GT manquées (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		
		poolé	—	—	—	15 en baisse / 10 en hausse / 1170 inchangés	1195	306	310
clean	Secondaire — lésions faux positifs	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		
		poolé	—	—	—	11 en baisse / 19 en hausse / 1165 inchangés	1195	72	57

ne sont pas de même nature : H1 était une affirmation directionnelle sur un mécanisme et elle est contredite dans son mécanisme ; H2 était une affirmation d'équivalence et elle échoue parce que l'effet est réel, quoique petit.

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: Le critère primaire pré-enregistré, apparié par patient, dans les deux régimes de décodage. À gauche : les lésions GT manquées entièrement (Dice = 0), le critère figé ; à droite : les faux positifs lésionnels. Le régime brut montre les deux clauses de H1 échouer en sens opposés ; le régime clean montre le filtrage officiel de composantes absorber l'essentiel des deux.

4.2 Endpoint primaire : un déficit constant en signe et réel en fréquence

Le Tableau 3 donne l'endpoint primaire déclaré — Dice et HD95 lésion-wise officiels, moyenne des trois régions par patient, régime `clean`, poolé sur cinq folds, n = 1195, apparié contre la baseline du même fold et de la même graine.

Le delta sur la moyenne des régions est **-0,00376** (IC 95 % [-0,00648 ; -0,00103], p de Holm = 8,0e-14), soit 1,25 fois le SESOI pré-spécifié, sur un Dice absolu de 0,8776 pour la baseline contre 0,8738 pour le bras blob. Par région, le signe ne change jamais : WT -0,00408 (IC [-0,00764 ; -0,00052], p de Holm = 2,5e-04), TC -0,00414 (IC [-0,00850 ; 0,00022], p = 1,6e-11), ET -0,00305 (IC [-0,00685 ; 0,00076], p = 2,3e-10). Le HD95 lésion-wise bouge dans le même sens sans atteindre la significativité sur la moyenne (+0,7772 mm, p de Holm = 1), l'ET étant la seule région significative (+0,6362 mm, p = 0,0283) — le Dice et le HD95 n'ont pas à être d'accord, et ici ils le sont en signe mais pas en force.

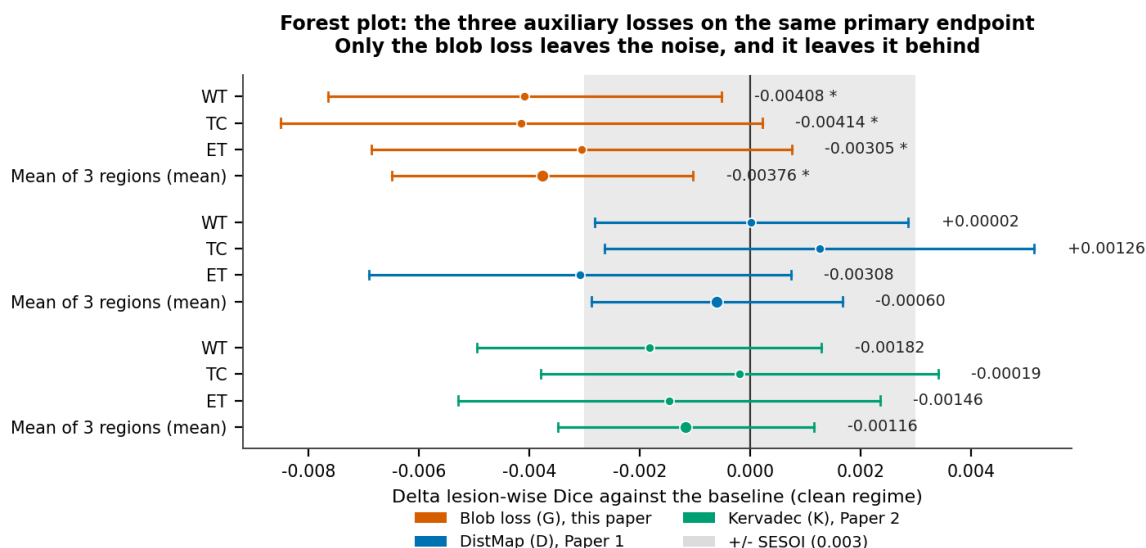
Lus avec les deux instruments : l'intervalle exclut zéro, donc le déficit est réel ; 480 cas sur 1195 sont meilleurs sous le bras blob, et sur les 1193 cas dont le score a bougé, 713 ont baissé ; mais la MDE propre à la moyenne appariée (0,00390) dépasse l'effet qu'elle détecte, donc l'**ampleur** reste incertaine entre « négligeable » et « le double du SESOI ». L'énoncé honnête est celui du §3.4, et c'est celui que porte le résumé.

Lecture. Le déficit sur la région moyenne est de -0.00376 (IC [-0.00648, -0.00103], p de Holm = 8.0e-14) : l'intervalle **exclut zéro**, le déficit est donc réel en fréquence, et il vaut 1.25× le SESOI pré-spécifié (0.003). L'IC 95 % exclut zéro pour 2 des 4 lignes rapportées (WT, Moyenne des 3 régions) ; les autres intervalles contiennent zéro et sont rapportés tels quels. Mais la moyenne appariée est ici un instrument émoussé : sd apparié 0.0481 pour un delta de 0.0038, d'où une MDE nominale 0.00390 > |delta|. Ce que le test de rangs dit, ce n'est pas « bien pire » mais « pire plus souvent que meilleur » : 713 des 1193 cas qui bougent baissent. Verdict publiable : **un déficit faible, constant, réel en fréquence, dont l'ampleur reste autour du seuil d'intérêt clinique.**

Tableau 3 — Endpoint primaire : CV poolée 5 folds, nettoyage officiel

Population : **n = 1195** cas held-out (5 folds, le cas dégénéré BraTS-GLI-00017-001 à vérité terrain vide retiré — la même population sur laquelle tous les autres bras de ce programme sont scorés). Métriques officielles BraTS-2023 **lésion-wise**, régime **clean** (nettoyage de composantes connexes par région à 1000/250/500 voxels). Wilcoxon signé apparié par patient, correction de Holm sur la famille {G, D, K} x {WT, TC, ET, moyenne}. Bras G = blob loss (beta = 0.5), référence B = baseline MedNeXt-B, mêmes plans, même fold, même graine.

Région	Dice B	Dice G	Delta Dice	IC 95 %	p (Holm)	Cas où G gagne	MDE	Delta HD95 (mm)	IC 95 %	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
Moyenne des 3 régions	0.8776	0.8738	-0.00376	[-0.00648, -0.00103]	8.0e-14	480/1195	0.00390	+0.777	[-0.457, 2.011]	1



* 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: La forêt de l’endpoint primaire : les trois termes auxiliaires contre la baseline en Dice lésion-wise, avec intervalles de confiance à 95 %, p-valeurs corrigées de Holm et le SESOI pré-spécifié dessiné en bande. Le bras blob est le seul des trois dont l’intervalle exclut zéro, et il l’exclut du mauvais côté.

4.3 Trois termes auxiliaires, un endpoint : seule la blob sort du bruit

Le Tableau 4 place les deux autres termes du programme sur la population, le régime et la famille de Holm identiques. La régression de distance signée donne -0,00060 (IC [-0,00287 ; 0,00167], p de Holm = 1) et la boundary loss -0,00116 (IC [-0,00348 ; 0,00116], p = 0,9547) : aucun des deux intervalles n’exclut zéro, et aucun n’est distinguable de la baseline. La blob donne -0,00376 avec un intervalle qui l’exclut.

Cette phrase ne doit pas être lue comme un classement de mérite. À l’intérieur de l’unique famille de Holm de 12 tests — 3 bras par 4 agrégations —, le terme qui sort du bruit est celui dont le **signe** est constant, pas celui dont l’effet est grand : les deux bras nuls sont nuls au sens strict d’un intervalle chevauchant zéro, et la blob est négative au sens strict

d'un intervalle entièrement en dessous. Les trois restent sous ou autour du SESOI en ampleur, et aucun n'est candidat au déploiement. Ce que la comparaison établit, c'est que « aligné sur la métrique lésion-wise » n'est pas la même propriété que « utile sur la métrique lésion-wise » — le seul terme construit pour imiter la métrique de classement est aussi le seul mesurablement pire sous elle.

Tableau 4 — Les trois loss auxiliaires sur le même endpoint primaire

Même population, même régime, même famille de Holm. DistMap (D, papier 1) et Kervadec (K, papier 2) sont donnés comme comparateurs : aucune des deux ne sort du bruit. La blob loss est la **seule** des trois dont l'intervalle de confiance exclut zéro — et il l'exclut du mauvais côté.

Loss auxiliaire	Delta Dice (moyenne)	IC 95 %	p (Holm)	Delta HD95 (mm)	p (Holm)	Cas gagnés / n	Verdict vs SESOI
Blob loss (G)	-0.00376	[-0.00648, -0.00103]	8.0e-14	+0.777	1	480/1195	IC excluant zéro - la seule des trois à sortir du bruit, et elle en sort par le bas
DistMap (D)	-0.00060	[-0.00287, 0.00167]	1	-0.079	0.1723	612/1195	IC contenant zéro - indiscernable de la baseline
Kervadec (K)	-0.00116	[-0.00348, 0.00116]	0.9547	+0.464	1	643/1195	IC contenant zéro - indiscernable de la baseline

4.4 Brut contre clean : le post-traitement est l'étage qui décide

Le Tableau 5 et la Fig. 3 mettent les deux régimes de décodage côte à côte. En régime brut le bras blob est à -0,0554 en Dice lésion-wise (p de Holm = $1,3e-46$) et +20,6579 mm en HD95 lésion-wise (p = $1,7e-27$). En régime clean cela devient -0,00376 (p = $8,0e-14$) et +0,7772 mm (p = 1). Le filtrage officiel de composantes absorbe donc **93,2 %** de l'écart de Dice et **96,2 %** de l'écart de HD95, et l'absorption est uniforme entre régions — de 92,4 % à 93,8 % de l'écart de Dice et de 95,1 % à 97,1 % de l'écart de HD95, les deltas par région allant de -0,0663 à -0,00408 sur WT, de -0,0543 à -0,00414 sur TC et de -0,0457 à -0,00305 sur ET.

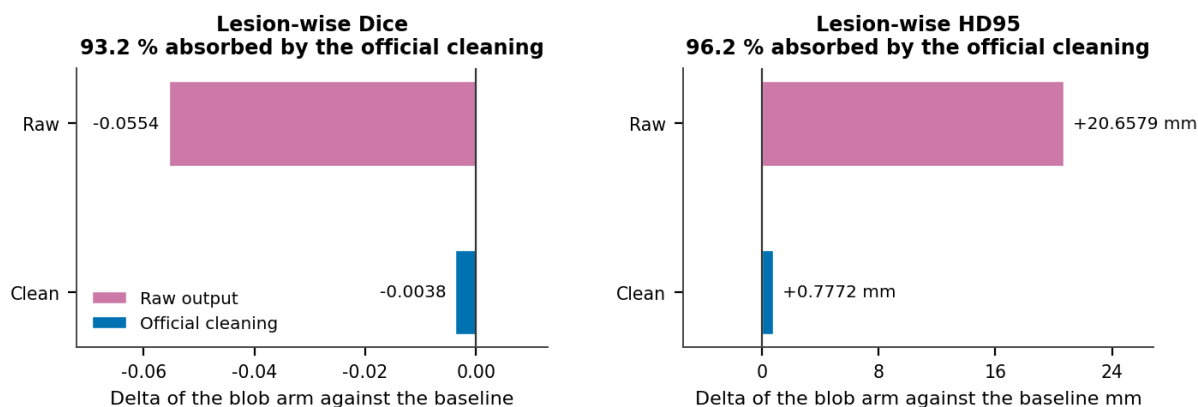
Deux conséquences en découlent, qui tirent dans des directions différentes. Sur le plan méthodologique, un papier qui ne rapporterait que le régime brut annoncerait un effondrement que le protocole du challenge ne mesure pas, et un papier qui ne rapporterait que le régime clean dissimulerait que son bras modifie assez la distribution de tailles des composantes pour qu'un filtre ait de l'effet. Les deux régimes sont donc dans chaque tableau susceptible de les porter, avec la part absorbée **calculée** et non assertée. Sur le plan substantiel, l'absorption est le mécanisme : le terme blob ajoute des petites composantes — dont certaines sont des lésions GT récupérées, dont la plupart non — et le filtre officiel supprime les petites composantes sans discernement. Ce qui survit est le résidu, et le résidu est un faible déficit.

Tableau 5 — Brut vs clean : ce que le post-traitement officiel absorbe

Le régime raw n'applique aucun nettoyage de composantes connexes ; clean applique les seuils officiels par région. La blob loss est conçue pour récupérer les petites instances, et en régime brut elle fait visiblement ce qu'elle annonce : elle détecte davantage de lésions. Le décodage officiel retire ensuite l'essentiel de ce qu'elle a ajouté, et le signal survivant est un faible déficit.

Métrique	Région	Delta brut	p (Holm) brut	Delta clean	p (Holm) clean	Part absorbée par le nettoyage officiel
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	Moyenne des 3 régions	-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	Moyenne des 3 régions	+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: Ce que le post-traitement officiel absorbe. Chaque paire de barres est le même écart bras-métrique avant et après nettoyage de composantes ; la part annotée est calculée depuis les deux, non citée. Publier en régime brut aurait annoncé un effondrement ; publier en clean seul aurait caché le mécanisme.

4.5 Le mécanisme en comptes

Les Tableaux 6a et 6b énoncent le mécanisme sans ratio. Sur 4007 lésions GT dans la validation croisée poolée, le régime **brut** donne au bras blob 3878 détections contre 3817 pour la baseline — un taux de détection de 0,9678 contre 0,9526 — au prix de 2029 composantes prédites non appariées contre 1418, soit une hausse de 43 %. C’est toute l’histoire de la loss en une ligne : elle trouve davantage de ce qui est là et invente davantage de ce qui n’y est pas.

Après nettoyage officiel, l’avantage de détection a pour l’essentiel disparu — 306 lésions manquées contre 310, taux de détection 0,9236 contre 0,9226 — tandis que l’écart de faux positifs se resserre sans se refermer (72 contre 57). Pour comparaison, les deux autres bras auxiliaires se comportent comme leurs papiers le décrivent : la régression de distance signée récupère une partie du rappel pour un coût modéré en faux positifs, et la boundary loss est la seule des trois à **réduire** les faux positifs en régime brut.

La ligne veto des mêmes tableaux anticipe le §4.8 : cc_B_G conserve le taux de détection de la baseline (0,9226) et retire des faux positifs (47 contre 57). C’est le seul endroit de ce papier où la contribution du bras blob est positive, et c’est une contribution d’**accord**, pas de segmentation.

Tableau 6a — Comptage lésionnel, régime brut (sans nettoyage)

Poolé sur les 5 folds. Voici le mécanisme, énoncé en comptages plutôt qu’en ratio : le bras blob trouve **davantage** de lésions GT que la baseline et le paie de **davantage** de composantes prédites non appariées.

Bras	Lésions GT	Détectées	Manquées	Taux de détection	Prédictions non appariées (FP)	FP par cas
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 veté par G	4007	3815	192	0.9521	903	0.7556
G veté par B	4007	3815	192	0.9521	886	0.7414
Vote majoritaire (2 sur 3)	4007	3824	183	0.9543	1190	0.9958

Tableau 6b — Comptage lésionnel, régime clean (nettoyage officiel)

Après le nettoyage officiel l'avantage de détection a disparu (les deux bras manquent presque le même nombre de lésions GT) tandis que l'écart de faux positifs se réduit sans se refermer. Le veto cc_B_G conserve le taux de détection de la baseline et retire des faux positifs.

Bras	Lésions GT	Détectées	Manquées	Taux de détection	Prédictions non appariées (FP)	FP par cas
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 veté par G	4007	3697	310	0.9226	47	0.0393
G veté par B	4007	3698	309	0.9229	57	0.0477
Vote majoritaire (2 sur 3)	4007	3685	322	0.9196	52	0.0435

4.6 Par fold : un signe constant sur cinq populations indépendantes

Le Tableau 7 décompose l'endpoint primaire fold par fold, chaque fold scoré contre la baseline du même fold, Holm sur 20 tests (5 folds × 4 régions), régime clean, métrique Dice lésion-wise. Les deltas clean sont -0,00216, -0,00561, -0,00425, -0,00301 et -0,00375 sur 239 cas chacun : **0 fold positif sur 5**, 5 négatifs, étendue 0,00346, moyenne des cinq -0,00376. Le régime brut est unanime aussi (0 fold positif), et son étendue vaut 4,65 fois celle du clean : 0,01607 contre 0,00346.

Les intervalles par fold contiennent presque tous zéro, et 1 fold seulement a un delta plus petit en valeur absolue que le SESOI — le signe est constant alors que la résolution par fold ne l'est pas. C'est précisément la situation qu'anticipait le garde-fou du pré-enregistrement : l'affirmation repose sur le test apparié **poolé** et sur la constance du signe sur cinq populations de patients disjointes, jamais sur un test par fold significatif. C'est aussi pourquoi le papier ne dit pas « la blob loss est significativement pire sur chaque fold » : il n'a pas la résolution pour le dire, et le dire serait la sur-affirmation que le SESOI existe pour empêcher.

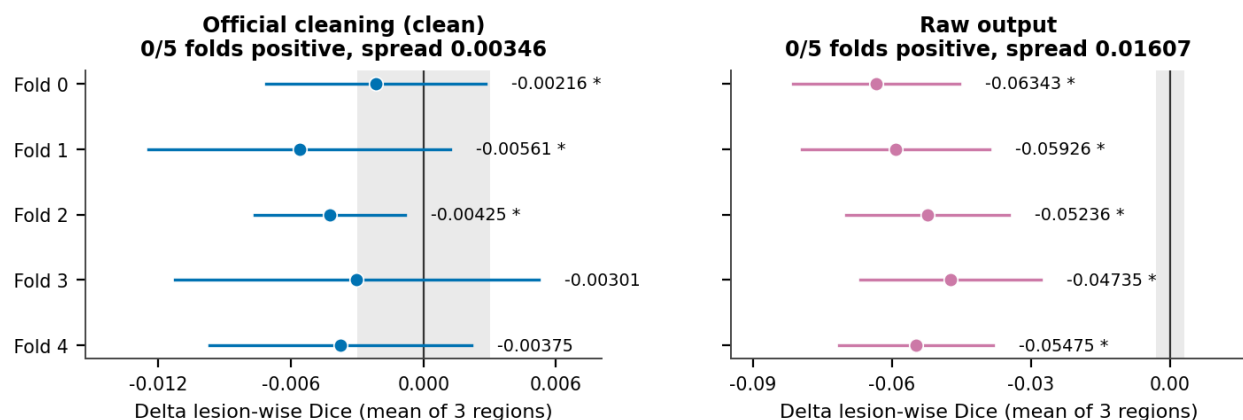
Tableau 7 — Décomposition par fold de l'endpoint primaire

Chaque fold est scoré séparément contre la baseline du même fold, puis poolé. Famille de Holm : 20 tests (5 folds x 4 régions) par régime. Le signe par fold est constant, ce qui rend l'intervalle poolé interprétable.

Fold	n	Delta Dice clean	IC 95 %	p (Holm)	Delta Dice brut	p (Holm) brut
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

Régime clean : **0/5 folds positifs**, 5 négatifs, étendue 0.00346, moyenne des 5 folds -0.00376, 1 fold sous le SESOI en valeur absolue. Régime brut : 0/5 folds positifs, étendue 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: Décomposition par fold de l'endpoint primaire dans les deux régimes de décodage, avec la bande du SESOI. Le signe est constant sur les cinq populations disjointes ; les intervalles par fold ne sont pas résolubles, ce qui est la raison pour laquelle la conclusion est pooling.

4.7 Le fold 0 seul : trois graines et un témoin de non-déterminisme

Le Tableau 8 est la raison pour laquelle aucune conclusion de ce papier ne repose sur un fold. Au fold 0, en régime clean, le bras blob à la graine 42 se tient à $-0,00081$ de la baseline (IC $[-0,00658 ; 0,00495]$, MDE $0,00824$) — pendant qu'un **second run de la baseline à graine identique** se tient à $+0,00134$. Les trois graines blob donnent $-0,00081$, $-0,00185$, $+0,00194$, s'étendant sur $0,00378$ autour d'un Dice lésion-wise moyen de $0,88140$: une étendue de $1,26$ fois le SESOI, plus grande que l'écart du fold 0 qu'elle est censée interpréter. Le verdict de l'évaluation elle-même est indiscernable (dans l'étendue des 3 graines), et ce papier l'adopte sans l'adoucir.

En régime brut le même fold est sans ambiguïté : $-0,05130$ pour le bras blob contre $+0,00390$ pour le témoin, verdict hors du bruit du témoin. Les deux régimes divergent sur le point de savoir si le fold 0 montre un effet, et la divergence est elle-même instructive — l'effet assez grand pour dépasser le bruit d'un fold seul est celui que le protocole officiel supprime ensuite.

Tableau 8 — Le fold 0 seul : trois graines et un témoin de non-déterminisme

Fold 0 seul, $n = 239$ après retrait du cas dégénéré. Le témoin est un **second run de la baseline à graine identique** : il mesure le non-déterminisme du GPU, pas une différence de modèle. Les trois graines du blob sont la réplication pré-spécifiée.

Bras	Delta Dice clean	IC 95 %	p (Holm)	Delta Dice brut	LW Dice du bras (clean)
Baseline, 2e run, graine identique (témoin)	$+0.00134$	$[-0.00223, 0.00492]$	0.3225	$+0.00390$	0.8830
Blob loss, graine 42	-0.00081	$[-0.00658, 0.00495]$	$4.0e-05$	-0.05130	0.8808
Blob loss, graine 1337	-0.00185	$[-0.00827, 0.00457]$	$8.8e-05$	-0.04448	0.8798
Blob loss, graine 2024	$+0.00194$	$[-0.00278, 0.00666]$	$2.5e-04$	-0.05227	0.8836

Verdict en régime clean : **indiscernable (dans l'étendue des 3 graines)** (blob -0.00081 , témoin $+0.00134$, trois graines -0.00081 , -0.00185 , $+0.00194$). Verdict en régime brut : **hors du bruit du témoin** (blob -0.05130 , témoin $+0.00390$). Étendue des graines sur le Dice lésion-wise clean : **0.00378** autour de 0.88140 — **plus large que le SESOI (0.003)**, ce

qui est exactement la raison pour laquelle un écart au fold 0 seul ne peut pas conclure : l'endpoint est la CV poolée du Tableau 3.

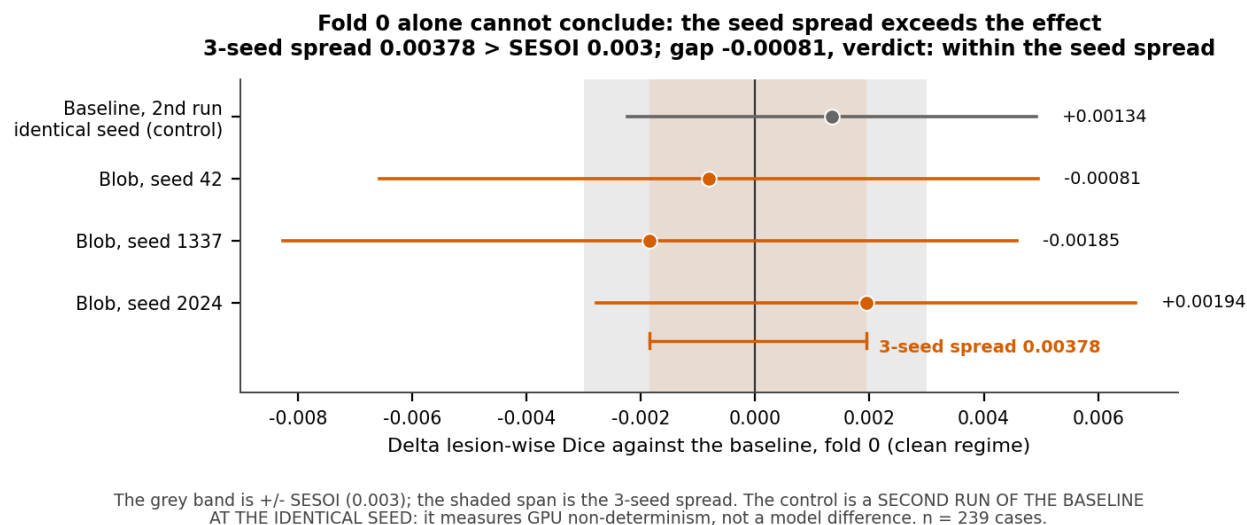


Figure 5: Le fold 0 comme instrument plutôt que comme résultat : les trois graines blob, le second run de la baseline à graine identique qui mesure le non-déterminisme GPU, et l'étendue des graines dessinée en plage contre le SESOI. L'étendue dépasse l'effet d'intérêt, donc aucun écart de fold 0 ne peut conclure.

4.8 Là où le bras paie : comme corroborateur dans un veto ordonné

Le Tableau 9 classe les douze vetos ordonnés à deux experts sur l'endpoint primaire, Holm sur les douze. cc_X_Y conserve les composantes connexes de X que Y corrobore. Le meilleur est cc_B_G — les composantes de la baseline, vetées par le bras blob — à +0,001137 (IC 95 % [0,000098 ; 0,002175], p brute = 0,0173, p de Holm = 0,156, MDE 0,00148), rang 1 sur 12. Le dauphin est cc_B_K à +0,000703 (p de Holm = 0,222).

Trois choses bornent ce nombre, et le papier les énonce toutes les trois chaque fois qu'il énonce le nombre. Il est **sous le SESOI** (0,003), comme les 12 vetos du tableau. Il **ne survit pas à Holm** (p = 0,156). 5 des douze sont nominalement significatifs avant correction, et les 3 qui y survivent — cc_G_K , cc_G_B , cc_G_D — sont **tous négatifs**, tous trois des opérateurs qui conservent les composantes du bras blob lui-même ; aucun delta positif de ce tableau ne survit à Holm. Et il est **asymétrique** : l'inversion de l'ordre donne cc_G_B à -0,002306, rang 11 (p de Holm = 3,9e-14). Seuls 4 des douze vetos sont positifs, et les quatre qui le sont conservent tous les composantes de la **baseline**.

Cette asymétrie est la partie interprétable. Un veto qui garde les composantes de X et demande à Y de les corroborer ne peut que retirer les faux positifs de X ; il ne peut jamais ajouter les découvertes de Y. Quand cc_B_G bat cc_G_B de la marge qu'il le fait, la mesure dit que les composantes du bras blob valent comme **vérification** de celles de la baseline et non comme composantes à **conserver**. Les comptes de composantes sont d'accord : après le veto, le taux de détection est celui de la baseline (0,9226 contre 0,9226) et les faux positifs tombent de 57 à 47. Le bras est un corroborateur. Sa valeur est celle d'un second avis qui coûte un run d'entraînement, ce qui est un mauvais marché — sauf si le run est déjà entraîné pour un autre usage, ce qui est le cas ordinaire dans un programme qui entraîne de toute façon un bras par terme auxiliaire et les score tous sur la même population.

cc_B_G (les composantes de la baseline vetées par le bras blob) est **rang 1 sur 12** à +0,001137, IC [0,000098, 0,002175], p de Holm = 0,1556 — **non significatif et sous le SESOI**, et il doit être rapporté comme tel. Le veto est **asymétrique** : inverser l'ordre (cc_G_B) le fait tomber au rang 11 à -0,002306. Ce qui survit est la composante de la baseline, pas celle du blob : le bras blob est utile comme **corroborateur**, pas comme **segmenteur**.

Tableau 9 — Là où le bras blob paie : comme membre d’un veto

Douze vetos ordonnés à 2 experts, même population et même régime, scorés sur la moyenne des trois régions. cc_X_Y conserve les composantes connexes de X que Y corrobore. Les rangs sont ceux du delta contre la baseline. Famille de Holm : les 12 vetos.

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

4.9 Reproductibilité et contrôles

La consolidation qui produit tous les nombres ci-dessus exécute 36/36 contrôles internes, tous verts, listés au Tableau 10 avec la valeur que chacun a vérifiée. Deux d’entre eux sont des rétro-validations contre un scoring antérieur et indépendant des MÊMES checkpoints par le harnais de validation croisée, assertées à $1e-9$: le delta poolé de ce bras, consigné là comme -0,003756015478, est recalculé ici avec un écart de $8,67e-19$, et le delta du veto cc_B_G , consigné comme 0,001136680289, avec un écart de $2,17e-19$. Ce papier ne peut donc pas raconter une autre histoire que cette lecture antérieure des mêmes artefacts — s’il essayait, un contrôle échouerait bruyamment au lieu d’échouer en silence.

5. Discussion

Le signe d’une loss par instance est une propriété du dataset. Le mécanisme de Kofler et al. — les faux positifs entrent au dénominateur de chaque instance, donc leur pénalité croît avec la multiplicité des instances — prédit un gain de précision là où les instances sont nombreuses et petites, et c’est ce que montre leur cohorte sclérose en plaques. Le gliome adulte est l’autre cas : une poignée de grosses composantes avec, autour, de petits satellites et des artefacts d’annotation. Là, faire compter chaque instance à égalité ne supprime pas les composantes parasites ; cela conduit le modèle à *tenter* chaque petite structure que la vérité terrain contient effectivement, et le prix est une hausse de 43 % des composantes non appariées. Nos comptes (1418 → 2029 en régime brut) et le benchmark nnU-Net indépendant de Bouteille et al. sur BraTS 2020 — précision au niveau instance 0,7159 contre 0,7708, F1 lésion-wise 0,4789 contre 0,4933, Dice volumétrique stable à 0,9168 contre 0,9171 (Bouteille et al. 2026) — concordent en direction à partir de données différentes, d’une configuration de backbone différente et d’un budget d’époques différent. Le résultat positif de Kundu et al. sur BraTS-METS 2025 (Dice foreground 0,615 contre une baseline à 0,59, cavité de résection de 0,150 à 0,221) se situe de l’autre côté de la même frontière : beaucoup de petites instances, positif en précision (Kundu et al. 2026). Trois datasets, deux signes, une seule variable explicative.

Le protocole d’évaluation n’est pas une enveloppe neutre. 93,2 % de l’écart de Dice et 96,2 % de l’écart de HD95 n’existent qu’en régime brut ; le filtrage officiel de composantes les retire. Un papier de méthode qui rapporte des prédictions brutes rapporte une grandeur qu’aucun challenge ne score, et un papier qui ne rapporte que des prédictions nettoyées ne peut pas montrer que sa méthode a modifié la distribution de tailles des composantes. C’est la version concrète et mesurée d’un reproche que la littérature des métriques fait en termes généraux (Maier-Hein et al. 2024 ; Reinke et al. 2024) : le post-traitement fait partie de la méthode, et le compte rendu honnête porte les deux régimes plus la part que l’un absorbe de l’autre.

Aligner la loss sur la métrique n’est pas l’optimiser. Le terme blob imite la structure par instance de la métrique de classement de plus près qu’aucun autre terme que nous ayons entraîné, et c’est le seul des trois mesurablement pire sous cette métrique. La raison est visible dans les comptes : la métrique score les composantes **appariées**, et elle compte un

Tableau 10 — Les contrôles exécutés par la consolidation elle-même

36/36 verts. Deux d’entre eux sont des rétro-validations contre un scoring antérieur et indépendant des *mêmes* checkpoints, à 1e-9 : ce papier ne peut pas raconter une autre histoire que cette lecture des mêmes artefacts. Les cinq identités ré-assertées sur chaque population et chaque régime sont regroupées ici en une ligne chacune, portant leur périmètre et leur verdict ; la valeur vérifiée par chaque run individuel est dans `analysis/blobloss_stats.json` (contrôles), qui est ce que le dépôt embarque.

Contrôle exécuté par la consolidation	Périmètre	Verdict	Détail
Fold 0 : 7200 mesures (5 bras × 2 régimes × 3 régions × 240 cas)	—	OK	7200
CV poolée : 200928 lignes = 28 bras × 2 régimes × 3 régions × 1196 cas	—	OK	200928 lignes, attendu 200928
CV poolée : 239 prédictions aux folds 1-4, 240 au fold 0 (écart documenté)	—	OK	{0: 240, 1: 239, 2: 239, 3: 239, 4: 239}
CV poolée : 1196 cas poolés, folds disjoints	—	OK	{’tous’: 1196, ’sans_degenere’: 1195}
7 runs à 300 époques terminées (lus des training logs)	—	OK	les 7 = 300
beta = 0.5, lambda_dist = 0.0 sur les 7 runs (relus des debug.json)	—	OK	betas={0.5} lambdas={0.0}
Campagne close 7/7, 0 échec	—	OK	{’runs’: 7, ’terminees’: 7, ’echecs’: 0}
Rétro-validation : delta poolé de ce bras contre un scoring antérieur indépendant des mêmes checkpoints (1e-09)	—	OK	publié -0.003756015478 vs recalculé -0.003756015478 (écart 8.67e-19)
Rétro-validation : delta du veto cc_B_G contre ce même scoring antérieur (1e-09)	—	OK	publié +0.001136680289 vs recalculé +0.001136680289 (écart 2.17e-19)
5 folds à 300 époques, tels que consignés par la campagne	—	OK	{’0’: 300, ’1’: 300, ’2’: 300, ’3’: 300, ’4’: 300}
Fold 0, clean : verdict repris tel quel du summary.json	—	OK	indiscernable (dans l’étendue des 3 graines)
Fold 0, brut : verdict repris tel quel du summary.json	—	OK	hors du bruit du témoin
L’étendue des 3 graines dépasse le SESOI en clean (un écart fold-0 seul ne conclut rien)	—	OK	étendue 3 graines = 0.0038, SESOI = 0.003
SESOI 0.003 inscrit dans le JSON	—	OK	0.003
Famille de Holm du critère primaire = les 3 régions (pré-enregistrement §3)	4 runs : 2 populations × 2 régimes	4/4 OK	3 tests
num_fn apparié == lésions GT manquées du comptage poolé (bras G)	4 runs : 2 populations × 2 régimes	4/4 OK	apparié 306 vs comptage 306 · apparié 129 vs comptage 129
Même identité pour la référence B	4 runs : 2 populations × 2 régimes	4/4 OK	apparié 310 vs comptage 310 · apparié 190 vs comptage 190
Les comptes appariés couvrent la population (moins + plus + inchangés == n)	4 runs : 2 populations × 2 régimes	4/4 OK	15 + 10 + 1171 vs n = 1196 · 37 + 2 + 1157 vs n = 1196 · ... (4 runs, tous listés dans le JSON)
num_fp apparié == faux positifs du comptage poolé (bras G)	4 runs : 2 populations × 2 régimes	4/4 OK	apparié 73 vs comptage 73 · apparié 2030 vs comptage 2030 · ... (4 runs, tous listés dans le JSON)
Critère primaire (compte apparié de lésions à Dice = 0) calculé dans les deux régimes	2 runs : les deux populations	2/2 OK	[’clean’, ’raw’]

Le md5 de chaque artefact lu par la consolidation est inscrit dans `analysis/blobloss_stats.json` et listé fichier par fichier dans le MANIFEST.sha256 du dépôt : tout lecteur peut ainsi vérifier que les nombres ci-dessus ont bien été calculés depuis les artefacts annoncés. Ce n’est pas reproduit en tableau ici — une liste d’empreintes est une propriété du dépôt, pas un argument du papier.

faux positif comme une lésion manquée entière. Une loss qui égalise le poids de chaque instance égalise donc aussi le poids de chaque instance **parasite**, et sur un dataset où la vérité terrain a peu de composantes, celles que le modèle produit en plus ne sont pas absorbées par la moyenne. L’alignement sur la métrique transfère les asymétries de la métrique dans le gradient — il n’hérite pas de son classement.

Ce qui aurait rendu ce bras utile. Deux choses que ce papier ne peut pas tester. D’abord le poids : $\beta = 1,0$ a été annoncé et jamais entraîné, et le compte de faux positifs en régime brut est exactement la grandeur qu’un β plus petit échangerait contre du rappel — le déficit pourrait donc être un déficit de β plutôt qu’un déficit de mécanisme. Ensuite le décodage : un filtre conditionné à la taille — qui ne conserve les petites composantes que là où le bras blob et la baseline sont d’accord — est la manière naturelle de dépenser l’asymétrie mesurée au §4.8, et `cc_B_G` en est la première approximation déterministe à +0,001137, sous le SESOI. Ni l’un ni l’autre n’est une affirmation ; ce sont les deux expériences que ce résultat implique, et toutes deux sont déclarées non testées ici.

Pourquoi le publier. La famille des loss auxiliaires mesurée sur ce backbone contient à ce jour deux nuls (papiers 1 et 2) et une réfutation directionnelle pré-enregistrée (celui-ci). Seul le dernier porte une prédiction figée avant les données et contredite par elles, ce qui en fait le seul capable d’échouer de façon informative (Varoquaux et al. 2022 ; Berger 1982). L’instruction pratique qu’il laisse est étroite et utilisable : sur le gliome adulte sous protocole officiel BraTS-2023, n’ajoutez pas de terme blob à Dice + CE pour récupérer des petites lésions — il les récupère et perd plus qu’il ne récupère ; si le bras est déjà entraîné, ses composantes valent pour **veto** de celles d’un autre.

6. Limites

Le poids de la loss n’est pas mesuré. Un seul β a été entraîné (0,5). Le bras $\beta = 1,0$ listé dans le pré-enregistrement n’a jamais été lancé, donc tout énoncé sur l’ampleur du déficit est conditionnel à ce poids unique. C’est la plus grande limite du papier, et c’est un choix de notre part, pas un accident : le bras a été dépriorité pendant la campagne et le pré-enregistrement n’a pas été amendé.

Un backbone, un dataset, une famille de décodage. MedNeXt-B sous nnU-Net v2 sur BraTS 2023 gliome adulte, avec les seuils officiels de composantes par région. Rien ici ne généralise à une autre architecture, à une autre pathologie — la littérature sur métastases et SEP pointe dans l’autre sens — ni à une autre politique de nettoyage.

Pas de stratification par taille de lésion en régime officiel. L’affirmation que le bras récupère des **petites** lésions est portée par les comptes (190 → 129 manquées en régime brut) mais pas par une ventilation par taille, parce que la seule distribution de tailles de lésions du dépôt vient du harnais post-hoc supprimé et est mesurée sur une population proxy (fit 180 / reserved 60). Elle est exclue plutôt que mélangée à un tableau à l’échelle officielle, ce qui veut dire que le mécanisme est **inféré** de totaux.

Pas de STAPLE, pas de combinaison apprise. La fusion probabiliste (Warfield et al. 2004) n’a aucune implémentation vérifiée dans ce dépôt, et les portes apprises sur des caches de probabilités d’experts en ont été supprimées le 2026-08-07 après avoir produit des classements que la métrique officielle n’a pas confirmés. Les douze opérateurs examinés ici sont déterministes et ordonnés, et leur famille est déclarée ; rien d’appris n’est scoré.

Les folds d’une validation croisée ne sont pas des échantillons indépendants. Pooler 1195 cas sur cinq folds les traite comme une seule population pour le test apparié, ce qui sous-estime la variance issue du découpage en folds (Nadeau et al. 2003). La décomposition par fold (Tableau 7) est rapportée pour cette raison : elle montre un signe constant et une étendue de 0,00346, et c’est l’affirmation la plus forte que le plan d’expérience soutient.

La moyenne appariée est grossière. MDE 0,00390 contre un delta de -0,00376 : le rejet vient des rangs signés, et un lecteur qui veut l’ampleur plutôt que la fréquence doit lire l’intervalle, pas la p-value. Réciproquement, 1 ligne du critère pré-enregistré est significative après Holm avec un intervalle contenant zéro — rapportée telle quelle plutôt qu’écartée.

Un artefact périmé a failli devenir une affirmation. Le `debug.json` du fold 0 rapporte `current_epoch = 15` là où le journal d’entraînement montre 300 époques terminées ; le compte d’époques du Tableau 1 est lu du journal seul, et la consolidation l’assure contre les comptes d’époques par fold consignés par les artefacts de la campagne d’entraînement elle-même. Tout nombre de ce papier qui pouvait venir de deux sources vient de celle qui a été vérifiée.

7. Conclusion

Une blob loss pré-enregistrée en aveugle, entraînée 300 époques sur cinq folds et trois graines du gliome adulte BraTS-2023, et scorée uniquement par la capsule officielle lésion-wise, ne fait pas ce que sa littérature source prédisait. On attendait d'elle qu'elle achète de la précision ; elle achète du rappel — 190 lésions GT manquées ramenées à 129 en régime brut, significatif dans 3/3 régions après Holm — et le paie de 2029 faux positifs lésionnels contre 1418, significatifs dans 3/3 régions. On attendait que son Dice lésion-wise soit neutre ; il vaut -0,00376 (IC [-0,00648 ; -0,00103], p de Holm = $8,0e-14$), soit 1,25 fois l'effet d'intérêt pré-spécifié, négatif sur 5 folds sur 5, et c'est le seul des trois termes auxiliaires du programme dont l'intervalle exclut zéro. Le post-traitement officiel absorbe 93,2 % de cet écart et 96,2 % de l'écart de HD95, ce qui est la raison pour laquelle les régimes brut et clean sont rapportés ensemble partout. La seule contribution positive du bras est comme corroborateur — le veto ordonné `cc_B_G` à +0,001137, rang 1 sur 12, sous le SESOI et non significatif après Holm, et asymétrique parce que ce qui survit à un veto est la composante du bras dont on conserve les composantes. L'énoncé généralisable porte sur la famille de loss plutôt que sur ce dataset : le signe d'une loss par instance suit la distribution de tailles d'instances des données, et aligner un objectif sur une métrique de classement transfère les asymétries de cette métrique dans le gradient au lieu d'hériter de son classement.

Contributions des auteurs

Guillaume Cassez — auteur principal : conception de l'étude, pré-enregistrement, entraînement des modèles, évaluations officielles, analyses statistiques, pipeline de consolidation et de contrôle, rédaction du manuscrit. **Stanislas Larnier** — formulation des questions de recherche, conseil méthodologique et relecture attentive des versions successives. Mêmes rôles que dans les papiers 1, 2 et 3.

Données, code et reproductibilité

Entraîné sur BraTS 2023 gliome adulte (données publiques de challenge). Chaque nombre de ce papier est produit par une chaîne déterministe qui ne lit que des artefacts commités : `scripts/paper5_collecter_chiffres.py` → `analysis/blobloss_stats.json` (36/36 contrôles, 16 sources avec md5) → `scripts/paper5_tables.py` → les quatre fichiers `table_paper5` dans `tables/` → `scripts/paper5_figures.py` → `figures/` avec `EM-PREINTES.json` → `scripts/paper5_manuscrit.py` → ce fichier et son miroir anglais, chaque marqueur étant tracé dans `tables/manuscrit_trace.json`. `scripts/paper5_coherence.py` relit toute la chaîne en mode fail-closed. Les métriques sont calculées par la capsule officielle BraTS-2023 (`external/BraTS-2023-Metrics`) et jamais par une réimplémentation locale. Aucune étape de ce pipeline n'exige de GPU.

Références

- 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, Stiefelbogen, 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.combiomed.2024.108944.