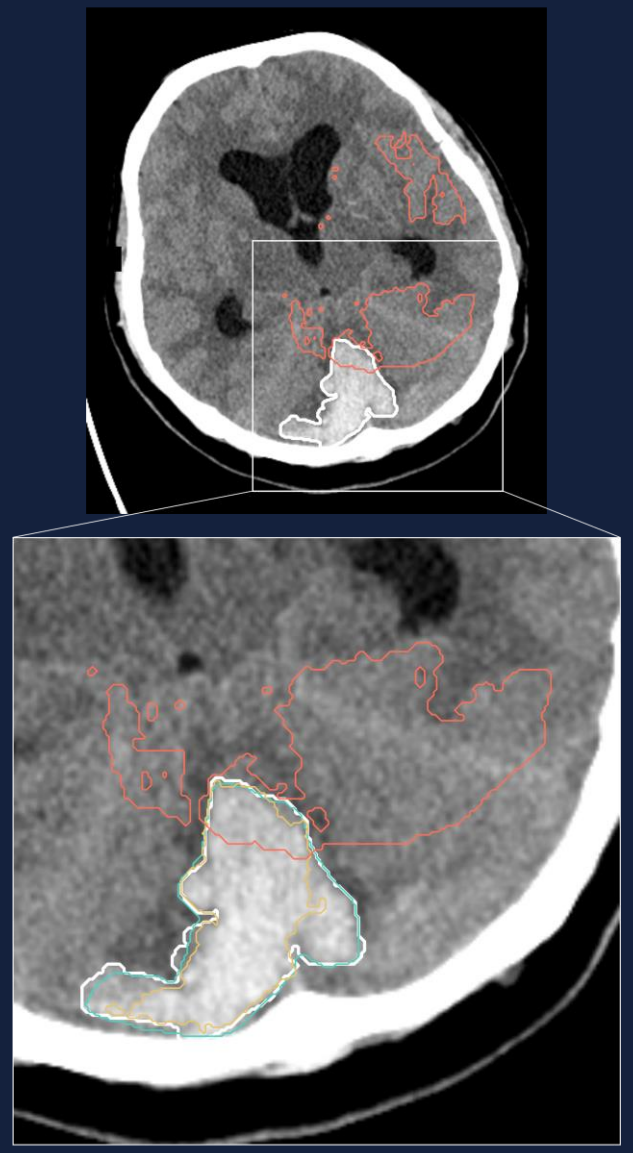
CONGRÈS ANNUEL DE LA SRQ · 13–15 NOVEMBRE 2026 · QUÉBEC

Segmentation automatisée de l'hémorragie intraparenchymateuse en TDM

Comparaison directe de sept outils publics préentraînés, appliqués sans réentraînement

Julien Ménès[1, 2] · **Laurent Létourneau-Guillon, MD, MSc, FRCPC**[2, 3, 4]

[1] Faculté de médecine, Université de Montréal
[2] Centre de recherche du CHUM (CRCHUM)
[3] Département de radiologie, radio-oncologie et médecine nucléaire, Université de Montréal
[4] Département de radiologie, Centre hospitalier de l'Université de Montréal (CHUM)



Exemple illustratif · coupe axiale

Financement : IRSC, subvention Projet 2025–2029 – Programme PREMIER UdeM

EN BREF

Sept outils, deux cohortes de validation : lequel segmente le mieux ?

7
Outils publics préentraînés

525
Examens analysés :
342 RSNA + 183 CHUM

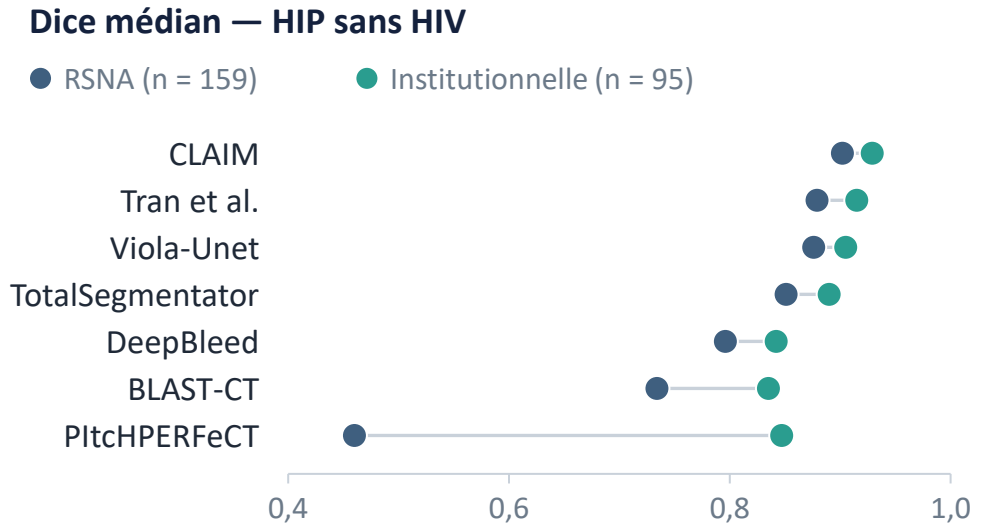
3 675
Segmentations évaluées

0,902 / 0,929
Dice médians de CLAIM — HIP sans HIV
RSNA (n = 159) / CHUM indépendant (n = 95)

Différences marquées de précision spatiale
RSNA sans HIV : Dice médians de 0,460 à 0,902 selon l'outil.

Mêmes outils en tête dans les deux cohortes
CLAIM, suivi de Tran et al., obtient le Dice médian le plus élevé.

Un bon accord volumétrique ne suffit pas
Un biais moyen quasi nul peut masquer de grandes erreurs individuelles.



Comparaison descriptive entre cohortes.

HIP : hémorragie intraparenchymateuse · HIV : hémorragie intraventriculaire

CONTEXTE CLINIQUE

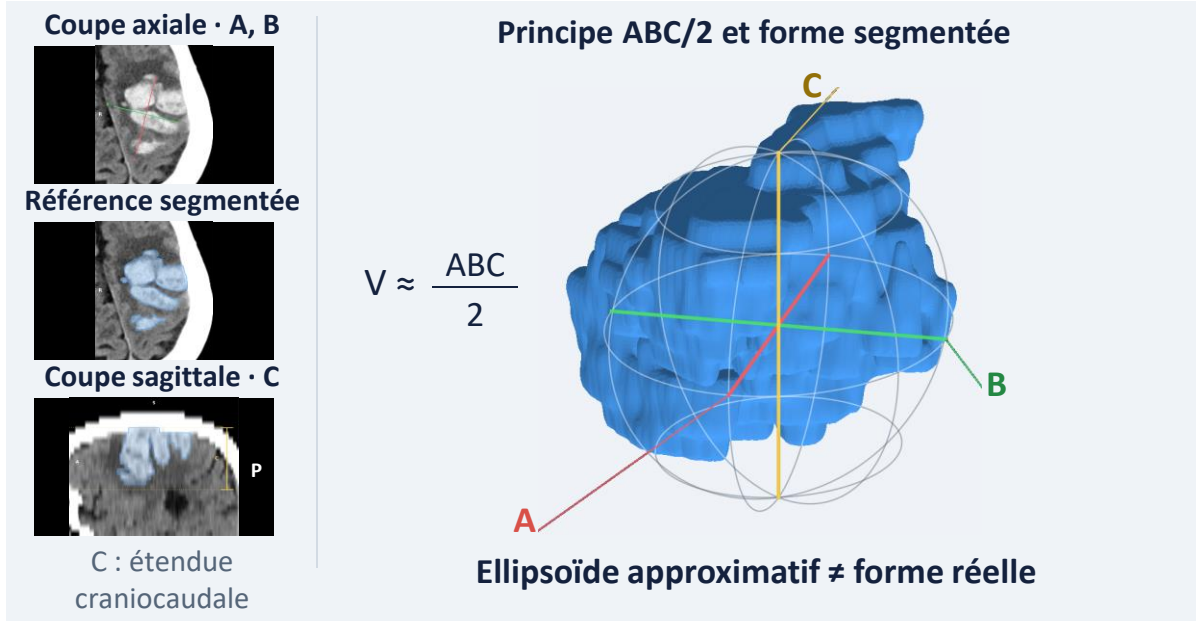
Le volume de l'HIP guide le pronostic et la prise en charge

Pronostic grave
L'HIP spontanée est l'une des formes d'AVC avec la mortalité la plus élevée et une morbidité importante chez les survivants (1).

Le volume initial prédit l'issue clinique
Prédicteur majeur du devenir fonctionnel, intégré au « score ICH » (seuil de 30 mL) (2).

Expansion de l'hématome
Augmentation > 33 % chez 20–40 % des patients examinés dans les 3 h suivant le début des symptômes : cible des thérapies anti-expansion (3).

Une mesure imparfaite
Segmentation manuelle précise mais chronophage. La formule ABC/2 couramment utilisée en clinique peut surestimer les hématomes irréguliers (4).



Pourquoi automatiser ?

Accélérer la quantification pour le pronostic, les essais anti-expansion, le suivi et la planification chirurgicale (5).

Réf. : (1) van Asch 2010; (2) Hemphill 2001; (3) Li 2024; (4) Huttner 2006; (5) Greenberg 2022.

PROBLÉMATIQUE ET OBJECTIF

Des outils publics hétérogènes, difficiles à comparer à partir des publications

Données différentes

Chaque outil est validé sur sa propre cohorte et par le même groupe – risque de surestimation de performance

Cibles anatomiques différentes

HIP seule, HIP + HIV, ou multiclasse : les sorties ne sont pas interchangeables.

Métriques différentes

Les mesures et références rapportées varient d'une publication à l'autre.

Compétitions d'IA ≠ déploiement clinique réel

Compétitions : soumissions optimisées. Ici : outils disponibles sans adaptation locale

Objectif

Comparer la précision de segmentation (coefficient de Dice) de sept outils publics préentraînés, appliqués tels que distribués, contre une référence experte commune, dans deux cohortes pour une application ciblant l'hémorragie intraparenchymateuse spontanée.

Hypothèse

Les outils diffèrent substantiellement en précision voxel par voxel.

Originalité de l'étude : comparer sept outils préentraînés, appliqués sans adaptation, sur les mêmes cohortes d'évaluation et contre la même référence experte indépendante.

26 outils candidats, 7 retenus

Recherche structurée, non systématique
Mai–juillet 2026 · PubMed, Google Scholar, GitHub, PyPI · compétitions d'IA
RSNA 2019, INSTANCE 2022, MBH-Seg 2024 et 2025

↓

26 outils candidats
Recherche d'outils publics de segmentation

↓

7 outils inclus
19 exclus, dont les outils commerciaux

Critères d'admissibilité

- Poids préentraînés ou application d'inférence proposés publiquement
- Inférence exécutable sans réentraînement ni ajustement
- Segmentation voxel par voxel incluant au moins l'HIP en TDM sans injection

Exclus : poids non publics, inférence non exécutable, absence de sortie HIP voxel par voxel, ou outils commerciaux (Qure.ai, Aidoc, RapidAI, Viz.ai…).

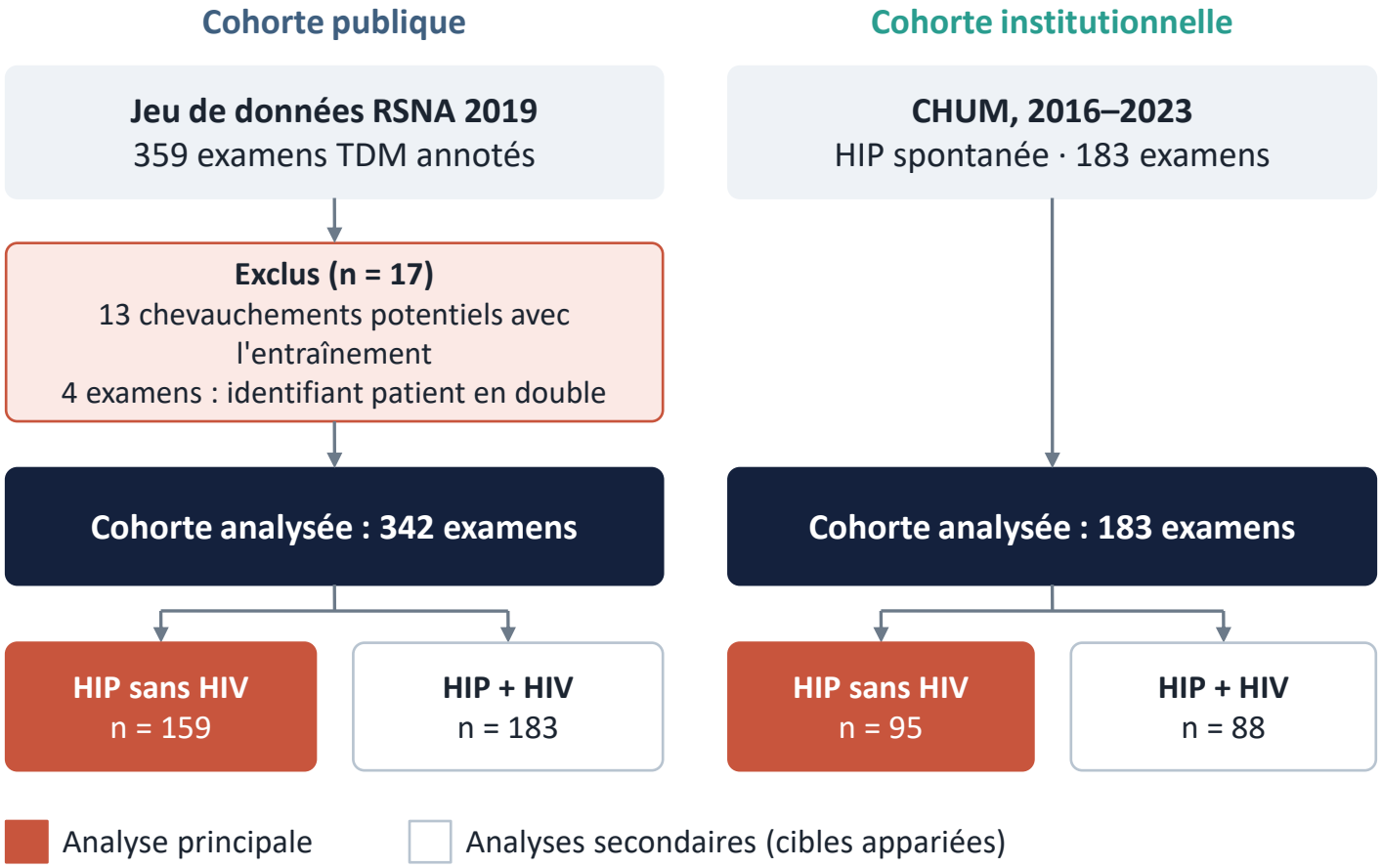
Sept outils aux origines et aux sorties hétérogènes

Outil	Architecture rapportée	Données de développement	Sortie native	Déploiement
CLAIM	Ensemble nnU-Net	MBH-Seg 2024 (BHSD, lignée RSNA)	Multiclasse (5 types)	Docker · GPU
Tran et al.	SwinUNETR + nnU-Net v2	ATACH-2 (HIP spontanée)	HIP + œdème	Application autonome (GUI)
Viola-Unet	U-Net à attention	INSTANCE 2022 + BHSD (lignée RSNA)	Multiclasse (5 types)	Docker · GPU
TotalSegmentator	nnU-Net (cerebral_bleed)	Jeu institutionnel dédié	Binaire (HIP + HIV)	Docker/pip · GPU
DeepBleed	CNN 3D (type V-Net)	MISTIE II/III (HIP spontanée)	Binaire (HIP + HIV)	Docker · GPU
BLAST-CT	DeepMedic (CNN 3D)	CENTER-TBI (traumatisme)	4 classes (TCC)	Docker · GPU
PItcHPERFeCT	Forêt aléatoire	MISTIE (HIP spontanée)	Binaire	Docker · CPU

Sept outils appliqués sans réentraînement ni ajustement des poids. Une sortie produite pour chacun des 525 examens.

BHSD : Brain Hemorrhage Segmentation Dataset; TCC : traumatisme craniocérébral. Réf. : Tran 2024; Liu 2023; Wasserthal 2023; Sharrock 2021; Monteiro 2020; Muschelli 2017.

525 examens : cohorte publique et cohorte institutionnelle



Cohorte publique RSNA 2019

- Sélection par un neuroradiologue parmi les examens avec HIP, surtout unifocale et d'allure spontanée
- Démographie, établissement et appareil non disponibles

Cohorte institutionnelle (CHUM)

- 111 H / 72 F; âge moyen 73,7 ans
- 94,5 % sur appareils Siemens
- Distincte des données de développement des sept outils

Référence : annotations 3D Slicer révisées par un neuroradiologue (12 ans d'expérience) (Wu et al., 2026). Approbation du comité d'éthique; consentement levé (données dépersonnalisées).

Analyse principale : même cible HIP, même évaluation

01 Préparation
DICOM → NIfTI standardisé
3D Slicer, correction
de l'inclinaison

02 Inférence
6 outils Docker (GPU, sauf
PItcHPERFeCT sur CPU)
Tran : appli. autonome

03 Harmonisation
HIP extraite des sorties
multiclasses ou HIP + œdème.
DeepBleed, TotalSegmentator :
sortie HIP + HIV, évaluée sur les
examens sans HIV.

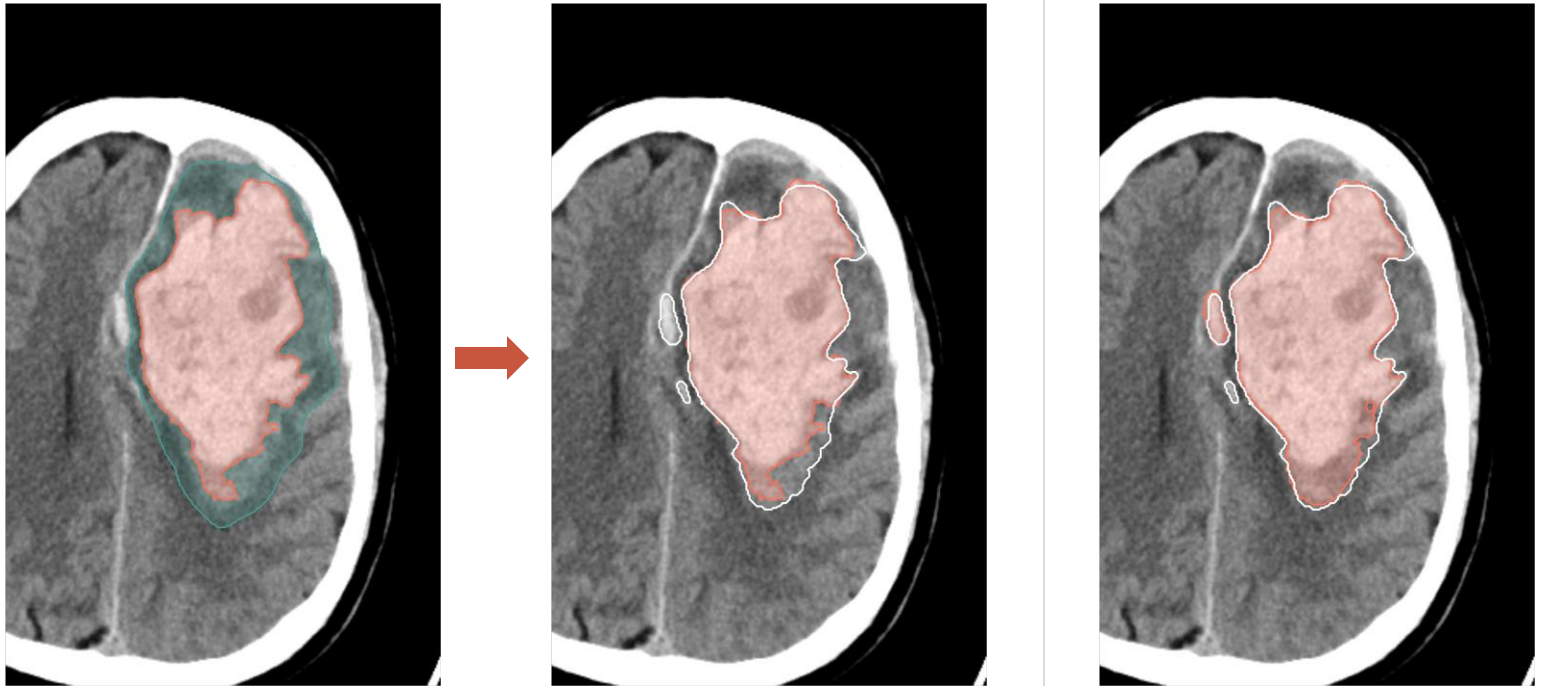
04 Évaluation
Coefficient de Dice : critère principal
Accord volumétrique : secondaire
Référence experte commune

Harmonisation des étiquettes

Tran et al.
Sortie native : HIP + œdème

Tran et al.
HIP retenue, œdème exclu

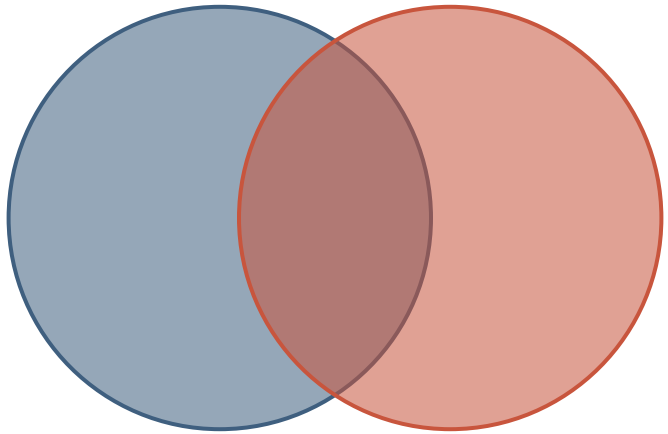
TotalSegmentator
Sortie binaire inchangée



Même référence HIP pour l'analyse principale sans HIV

HIP prédite · Œdème prédit · Référence HIP

Le coefficient de Dice comme critère principal



Dice = 2 |A ∩ B| / (|A| + |B|)

0 = aucun chevauchement

1 = superposition parfaite

Critère principal

- Dice par examen, y compris les prédictions vides (Dice = 0)
- Médiane (Q1–Q3), IC à 95 % par bootstrap BCa (10 000 tirages)
- Wilcoxon des rangs signés, correction de Holm (21 paires par cohorte)

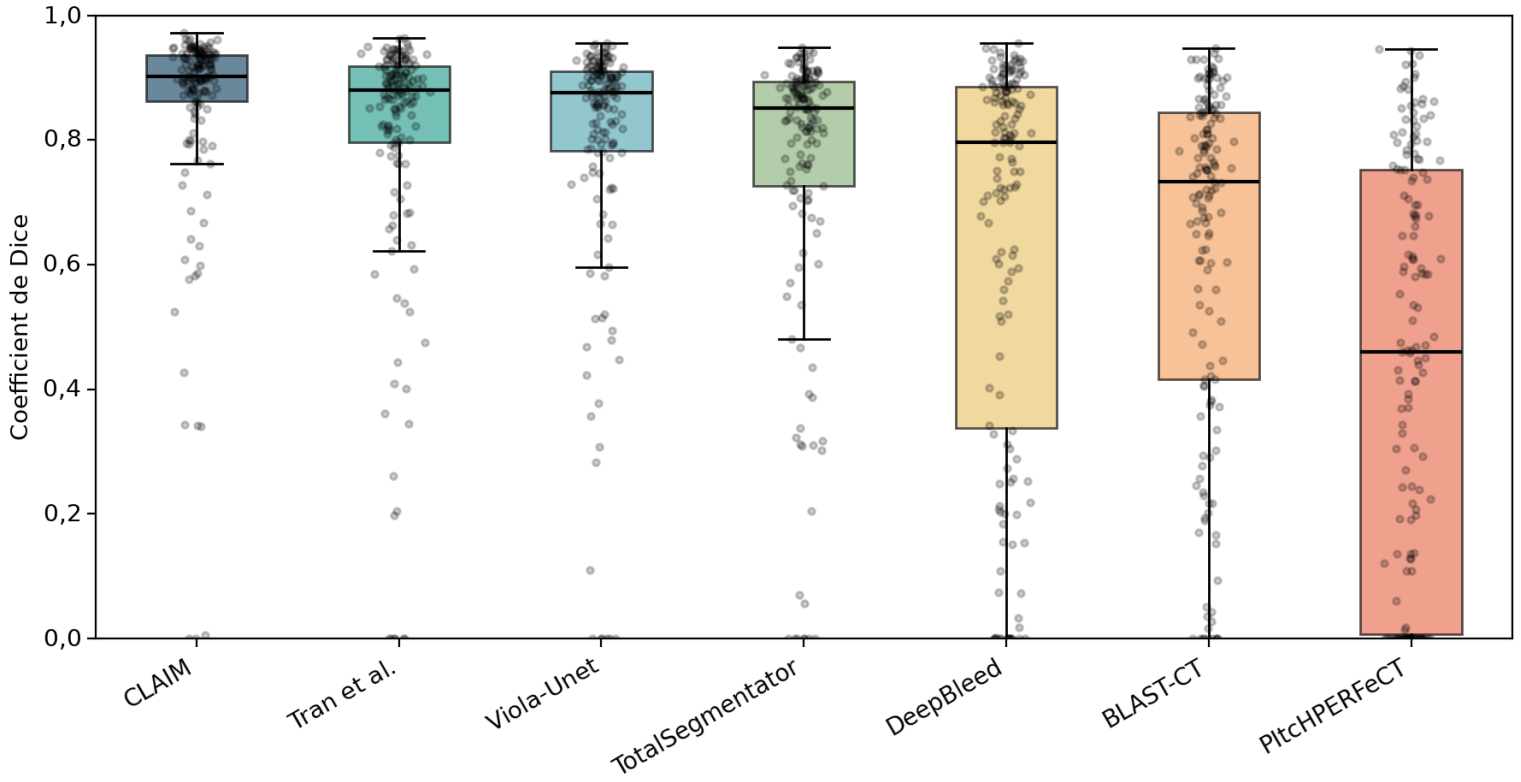
Analyses secondaires

- Accord volumétrique : coefficient de concordance (CCC) de Lin, diagramme de Bland–Altman
- Référence multilecteurs plutôt qu'un seul annotateur expert : consensus 2 sur 3 experts (30 examens)
- Cohortes avec HIV et cohortes totales, avec ou sans HIV

Analyse principale : HIP sans HIV. Comparaisons appariées au sein de chaque cohorte.

Q1–Q3 : quartiles; BCa : bootstrap corrigé du biais et accéléré; CCC : coefficient de corrélation de concordance.

CLAIM obtient le Dice médian le plus élevé



Points : examens · Boîtes : Q1–Q3 · Trait central : médiane

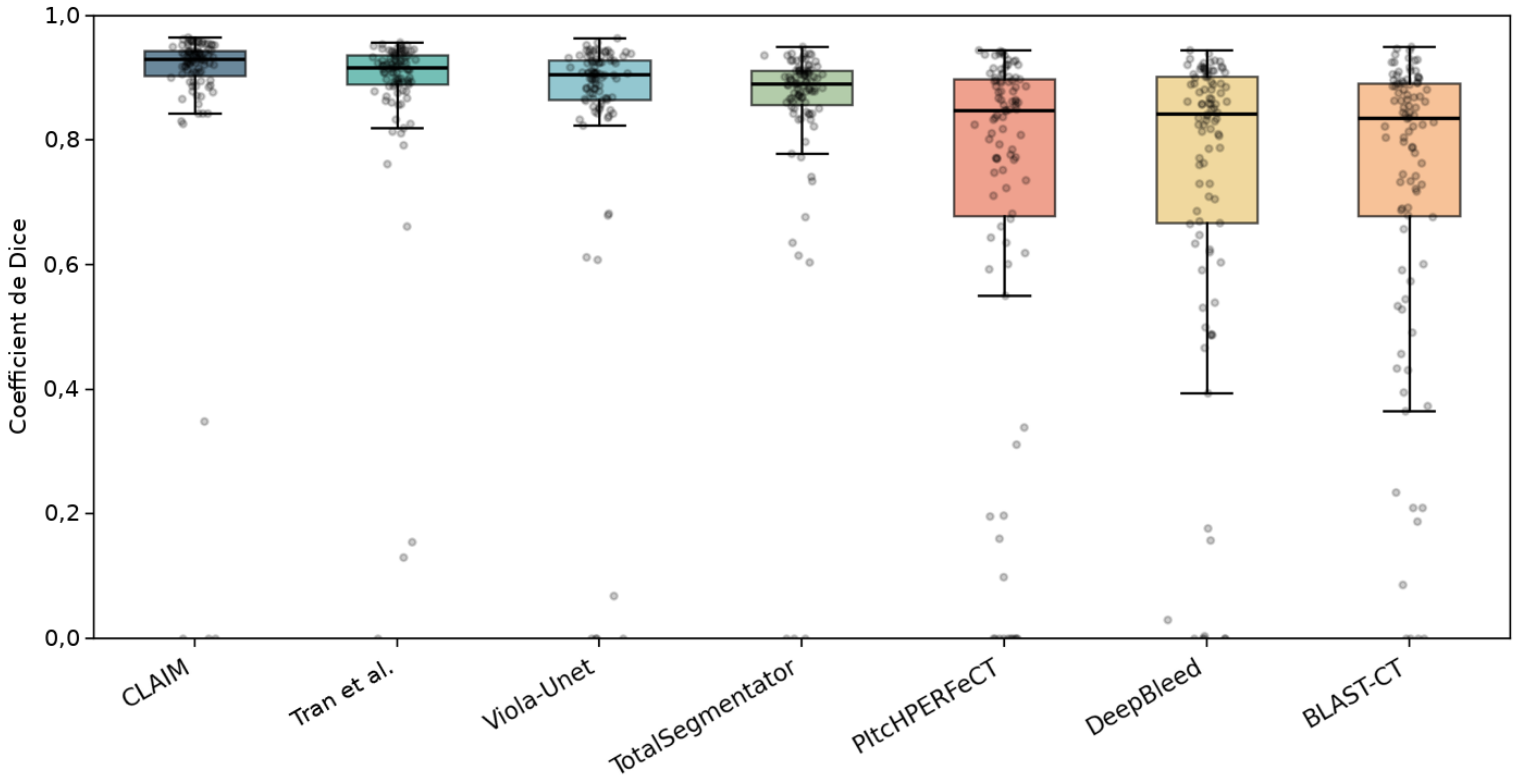
0,902

Dice médian de CLAIM
Q1–Q3 : 0,862–0,937

Tran et al.	**0,879**
Viola-Unet	**0,876**
TotalSegmentator	**0,851**
DeepBleed	**0,796**
BLAST-CT	**0,734**
PItcHPERFeCT	**0,460**

19/21 comparaisons significatives (Holm); CLAIM–Tran et al. : p ajusté < 0,001.
Non significatives : Tran et al.–Viola-Unet et DeepBleed–BLAST-CT.

Cohorte indépendante : CLAIM et Tran et al. restent en tête



0,929

Dice médian de CLAIM
Q1–Q3 : 0,903–0,944

Tran et al.	**0,915**
Viola-Unet	**0,905**
TotalSegmentator	**0,890**
PItcHPERFeCT	**0,847**
DeepBleed	**0,842**
BLAST-CT	**0,835**

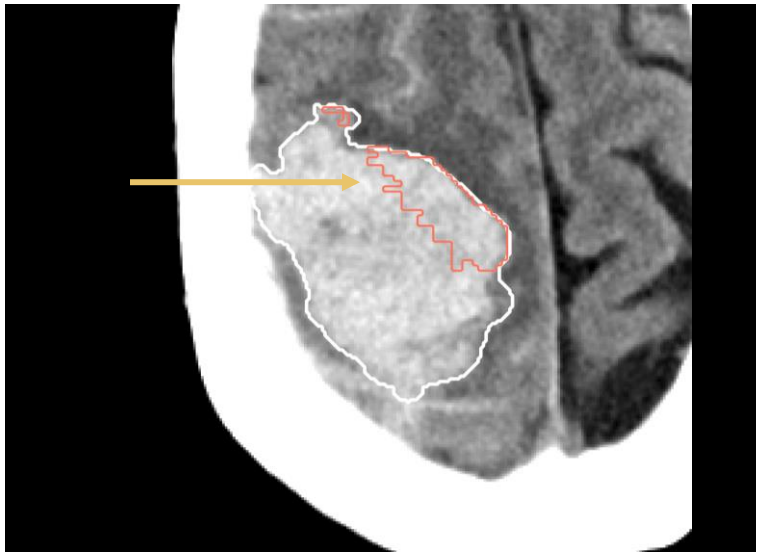
Points : examens · Boîtes : Q1–Q3 · Trait central : médiane

Écart CLAIM–Tran et al. faible mais significatif (p ajusté < 0,001); importance clinique à déterminer.
17/21 comparaisons significatives (Holm). Non significatives : Viola-Unet–TotalSegmentator et les trois comparaisons entre DeepBleed, BLAST-CT et PItcHPERFeCT.

Trois types d'erreurs de segmentation

Sous-segmentation

Une partie de l'hématome n'est pas segmentée

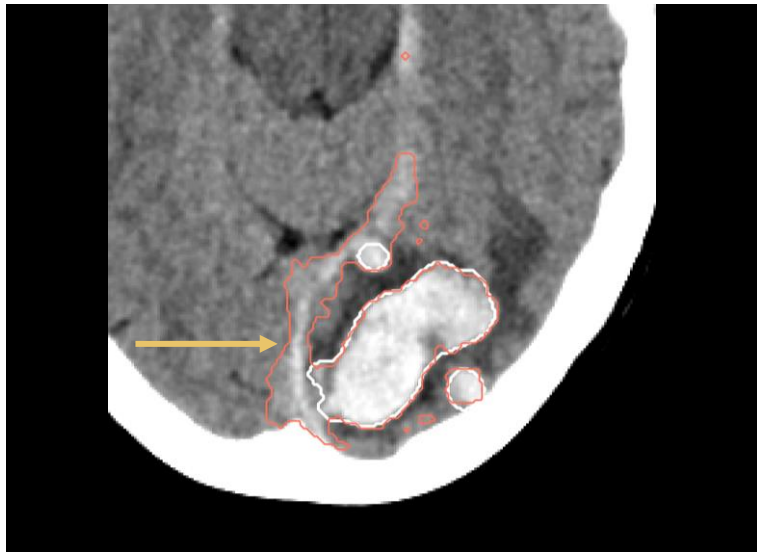


BLAST-CT · Dice 3D = 0,57

Volume réf./prédit : 39,0 / 16,0 mL

Sur-segmentation

La prédiction déborde de l'hématome (notamment sur la faux)



DeepBleed · Dice 3D = 0,60

Volume réf./prédit : 10,9 / 23,7 mL

Hématome sous-dural inclus

Faux positif pour l'HIP malgré un volume quasi identique



TotalSegmentator · Dice 3D = 0,83

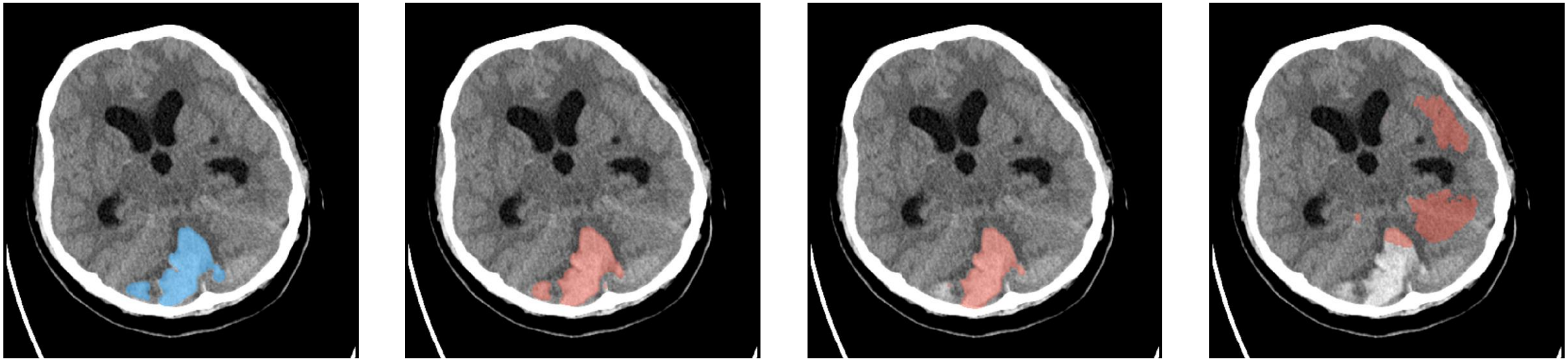
Volume réf./prédit : 103,6 / 105,0 mL

Référence experte — Prédiction

Le score résume la performance; l'examen du masque révèle la nature de l'erreur.

Exemples sélectionnés selon des critères quantitatifs, puis revus visuellement; ils ne représentent pas la fréquence des erreurs. Agrandissements différents; mesures sur le volume 3D complet.

Des sorties très différentes pour un même hématome



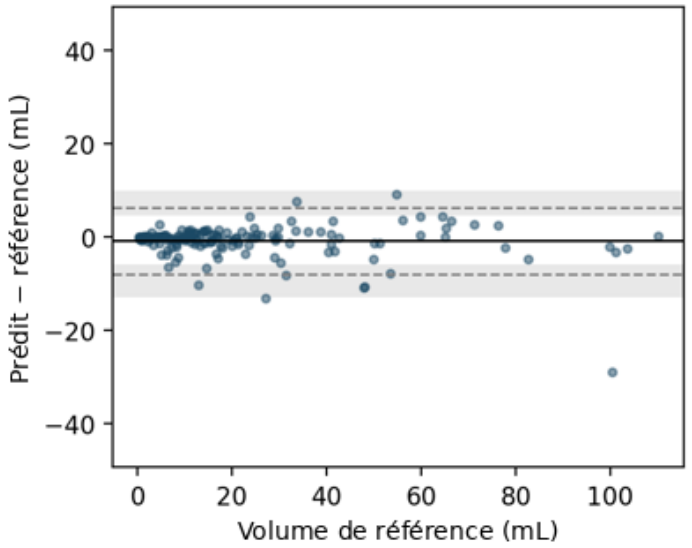
Référence experte	CLAIM	Tran et al.	DeepBleed
	Dice 3D = 0,94	Dice 3D = 0,91	Dice 3D = 0,03

HIP cérébelleuse avec hydrocéphalie obstructive. CLAIM et Tran et al. suivent étroitement la référence sur cette coupe. DeepBleed couvre peu l'hématome et segmente des régions hors cible. Dice calculé sur l'examen 3D complet.

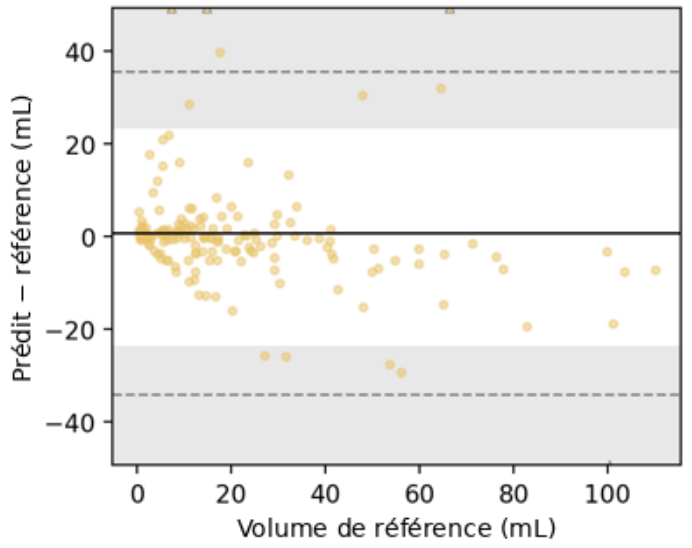
Bleu : référence; rouge : prédiction. Coupe axiale de plus grande surface d'HIP de référence; fenêtre cérébrale (L 80, N 40 UH). Autres outils sur ce même exemple : TotalSegmentator 0,90; BLAST-CT 0,89; Viola-Unet 0,88; PItcHPERFeCT 0,72.

Un biais moyen quasi nul peut masquer de grandes erreurs

CLAIM



DeepBleed



Deux outils illustratifs, axes identiques. Écart prédit − référence selon le volume de référence; pointillés : limites d'accord à 95 %. DeepBleed : 4 examens hors du cadre (+141, +83 et +57 mL; −80 mL), inclus dans le calcul du biais et des limites d'accord.

Biais proches de zéro, erreurs très différentes

	CLAIM	DeepBleed
Biais moyen	−0,8 mL	+0,8 mL
Limites d'accord (mL)	−8,0 à +6,3	−34,1 à +35,6
Dice médian	0,902	0,796

Corrélation ≠ accord : PItcHPERFeCT montre une corrélation linéaire élevée (r = 0,88) malgré le Dice médian le plus faible (0,460) et un biais de −10,0 mL.

Un volume global proche de la référence ne garantit pas une segmentation spatialement exacte.

CLAIM reste en tête contre un consensus de trois lecteurs

Neuroradiologue principal (nouvelle segmentation après 6 mois)

Second neuroradiologue (3 ans post-fellowship)

Neurologue vasculaire (10 ans post-fellowship)

Consensus
vote majoritaire
2 sur 3
voxel par voxel

Annotations indépendantes de novo, à l'aveugle des autres annotations et des sorties des modèles (Wu et al., 2026).

Analyse descriptive, sans test de comparaison, sur 30 examens institutionnels. CLAIM conserve le Dice médian le plus élevé; aucune équivalence à l'humain n'est établie.



Dice médian contre le consensus

Analyses secondaires : résultats selon la cible anatomique

Cible de référence selon la sortie

HIP séparable ou binaire HIP

- CLAIM
- Tran et al.
- Viola-Unet
- BLAST-CT
- PItcHPERFeCT

→ **Référence : HIP seule (HIV exclue)**

HIP + HIV combinées

- DeepBleed
- TotalSegmentator

→ **Référence : HIP + HIV**

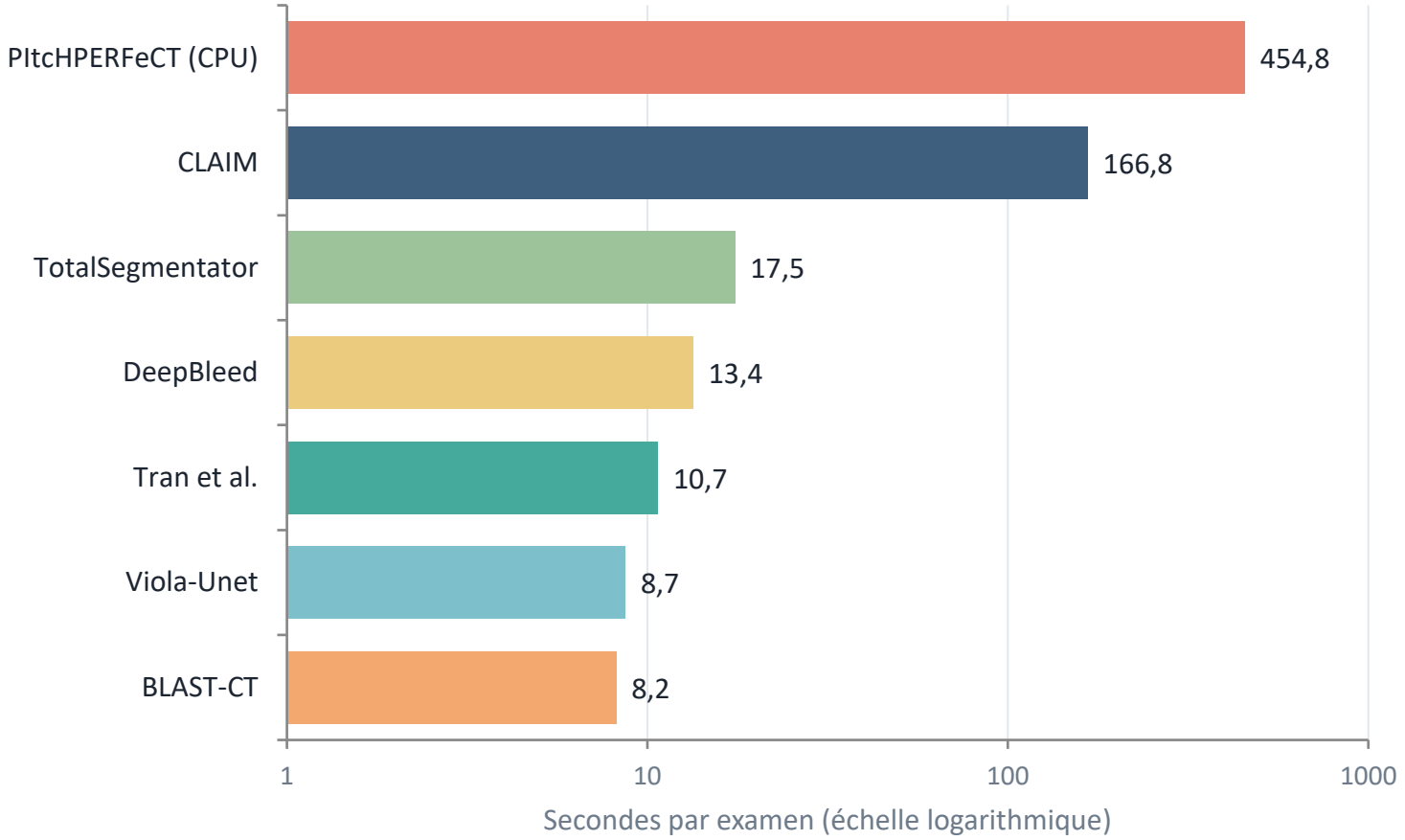
Cibles appariées selon l'outil : comparaison non strictement équivalente.

Outil	RSNA HIP + HIV (n = 183)	RSNA totale (n = 342)	CHUM HIP + HIV (n = 88)	CHUM totale (n = 183)
CLAIM	**0,915**	**0,913**	**0,913**	**0,923**
Tran et al.	**0,889**	**0,884**	**0,907**	**0,911**
Viola-Unet	0,857	0,867	0,880	0,897
TotalSegmentator †	0,869	0,864	0,887	0,889
DeepBleed †	0,824	0,812	0,856	0,852
BLAST-CT	0,678	0,708	0,775	0,797
PItcHPERFeCT	0,700	0,624	0,760	0,812

Dice médian par cohorte. Cohortes totales : HIP avec ou sans HIV. † Sortie binaire évaluée contre la cible HIP + HIV.

CLAIM et Tran et al. présentent les Dice médians les plus élevés selon les cibles retenues; celles-ci diffèrent entre les deux groupes d'outils.

Temps de traitement observés selon l'outil



CLAIM
Temps observé plus long (166,8 s par examen) dans les conditions de déploiement testées

Tran et al.
10,7 s par examen avec l'application autonome fournie par l'auteur; automatisation sans interface graphique non vérifiée.

Formats variés
Distribués en conteneur Docker, via pip, en application autonome ou en paquet R : l'intégration dépend de l'infrastructure locale.

Temps moyen sur 20 examens, initialisation incluse (GPU RTX 3090; PItcHPERFeCT sur CPU). Conditions hétérogènes : mesures descriptives, non standardisées.

La population d'entraînement n'explique pas tout

Tâche d'origine distincte

BLAST-CT a été développé pour le traumatisme craniocérébral (CENTER-TBI) : transfert vers une autre population clinique.

Même contexte, résultats différents

Tran et al., DeepBleed et PItcHPERFeCT proviennent tous d'essais sur l'HIP spontanée, mais leurs performances diffèrent fortement.

Cibles d’entraînement plus diverses

CLAIM et Viola-Unet, développés sur des jeux de données multiclasses non limités à l'HIP spontanée, figurent parmi les meilleurs.

Validation indépendante : CLAIM et Tran et al. restent en tête au CHUM, ce qui renforce la cohérence du résultat principal et minimise l’impact potentiel d’un chevauchement résiduel entre les données RSNA publiques d’évaluation et d’entraînement.

Contexte d’évaluation : les Dice médians sont plus élevés au CHUM, cohorte plus homogène d’HIP spontanées aiguës (< 24 h). La sélection RSNA a suréchantillonné des cas complexes (modifications chirurgicales, artéfacts), ce qui pourrait expliquer au moins en partie l’écart de performance.

Deux cohortes, des contextes d'évaluation différents

CHUM

HIP spontanée aiguë (< 24 h)



Hématome compact, bien délimité

RSNA

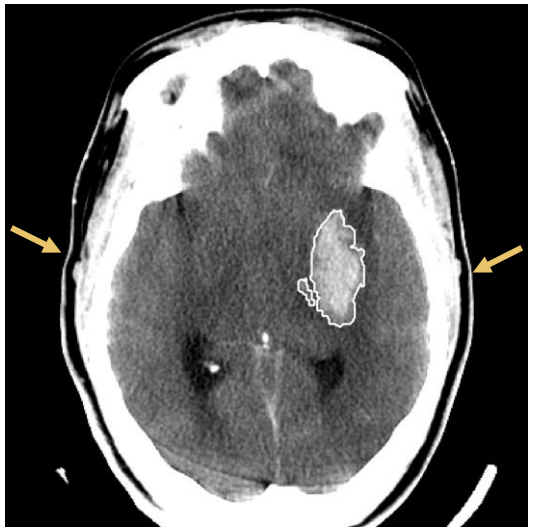
Modifications postopératoires



Air dans le lit opératoire

RSNA

Artéfacts de mouvement



Artéfacts légers : contours dédoublés (flèches)

RSNA

Artéfacts métalliques



Stries à travers l'hémisphère

La complexité des cas pourrait contribuer aux écarts de Dice entre cohortes; son effet propre n'a pas été mesuré.

Contour blanc : segmentation de référence. Exemples illustratifs, sans estimation de leur fréquence dans les cohortes analysées.

Une base pour choisir les outils à valider localement

1. Présélectionner

CLAIM et Tran et al. : meilleurs Dice médians dans les deux cohortes

2. Valider localement

Sur des cas institutionnels représentatifs de l'usage visé

3. Revoir les masques

Nature des erreurs et conséquences pour l'application

4. Intégrer

Ressources de calcul, compartiments produits, flux de travail

Usages potentiels et exigences de validation

Pronostic

Volume initial (score ICH) : de petites erreurs comptent près des seuils

Surveillance

Suivi de l'expansion de l'hématome sur des examens répétés

Essais et chirurgie

Seuils d'éligibilité (p. ex. 30 à 80 mL dans ENRICH) et planification

Radiomique / IA

Le masque sert d'entrée : la cartographie spatiale exacte est essentielle

Réf. : Hemphill 2001 (score ICH); Pradilla 2024 (ENRICH).

Limites à considérer

Examens positifs seulement

Spécificité non évaluée sur des examens sans HIP; faux positifs possibles même sur examens positifs.

Centre unique

Cohorte institutionnelle monocentrique, 94,5 % Siemens : généralisation à confirmer.

Devis rétrospectif

Cohorte RSNA sélectionnée et non consécutive; étiologies inconnues.

Sous-étude multilecteurs

30 examens institutionnels, limitée à l’HIP.

Analyses secondaires

Cohortes avec HIV : cibles de référence différentes selon l'outil.

Portée du Dice

Ne caractérise ni tous les modes d'erreur ni l'impact sur les décisions cliniques.

PERSPECTIVES

Prochaines étapes du projet

Revue qualitative des masques

Décrire les erreurs par outil.

Examens sans hémorragie

Quantifier les faux positifs.

Modèle multiclasse de l'équipe

Comparer sur les mêmes sous-groupes, avec une référence indépendante.

Impact clinique

Évaluer les conséquences des erreurs.

Diffusion

Manuscrit et partage du code.

CONTRIBUTION PERSONNELLE

Ma contribution : de la conception à l'évaluation comparative

Conception

- J'ai proposé l'idée de l'évaluation comparative au Dr Létourneau-Guillon
- J'ai recherché et sélectionné les outils admissibles
- J'ai conçu le cadre d'évaluation, affiné avec mon directeur

Réalisation

- J'ai installé, configuré et débogué les sept outils
- J'ai développé les scripts d'inférence, d'extraction des masques et de contrôle qualité
- J'ai recherché les chevauchements et constitué les ensembles d'analyse

Analyse et diffusion

- J'ai réalisé les analyses statistiques et l'évaluation contre le consensus
- J'ai préparé les tableaux, figures et documents supplémentaires
- J'ai rédigé le manuscrit initial, le rapport PREMIER et cette affiche

Encadrement : Dr L. Létourneau-Guillon (supervision scientifique, accès aux données, orientation méthodologique, relecture critique). Images et annotations de référence préexistantes de l'équipe.

CONCLUSION

Trois messages à retenir

1. **Les sept outils diffèrent en précision spatiale dans l’analyse principale HIP sans HIV.**
   Dice médians de 0,460 à 0,902 dans la cohorte RSNA

2. **CLAIM obtient le Dice médian le plus élevé dans les deux cohortes, suivi de Tran et al. La portée clinique de leur écart reste à établir.**
   RSNA : 0,902 vs 0,879 · Institutionnelle : 0,929 vs 0,915

3. **L'accord volumétrique complète l'évaluation spatiale, mais ne la remplace pas. Importance d’une vérification visuelle des masques.**
   Biais quasi nul ≠ segmentation exacte

Pour les centres : valider localement les meilleurs candidats et revoir les masques avant toute intégration en recherche ou en clinique.

Références choisies

1. van Asch CJ et al. Lancet Neurol. 2010;9(2):167-76.
2. Hemphill JC 3rd et al. The ICH score. Stroke. 2001;32(4):891-7.
3. Li Q et al. Code ICH: a call to action. Stroke. 2024;55(2):494-505.
4. Huttner HB et al. ABC/2 vs planimetric analysis. Stroke. 2006;37(2):404-8.
5. Greenberg SM et al. 2022 AHA/ASA ICH guideline. Stroke. 2022;53(7):e282-e361.
6. Flanders AE et al. RSNA 2019 Brain CT Hemorrhage Challenge. Radiol Artif Intell. 2020;2(3):e190211.
7. Wu AN et al. Multiclass segmentation of ICH, IVH and PHE. Radiology Advances. 2026.
8. Xie Y et al. Multi-class brain hemorrhage segmentation (MBH-Seg 2024). Zenodo; 2024.
9. Tran AT et al. Bioengineering (Basel). 2024;11(12):1274.
10. Liu Q et al. Viola-Unet. IEEE ISBI 2023:1-5.
11. Wasserthal J et al. TotalSegmentator. Radiol Artif Intell. 2023;5(5):e230024.
12. Sharrock MF et al. DeepBleed. Neuroinformatics. 2021;19(3):403-15.
13. Monteiro M et al. BLAST-CT. Lancet Digit Health. 2020;2(6):e314-e22.
14. Muschelli J et al. PItcHPERFeCT. Neuroimage Clin. 2017;14:379-90.
15. Pradilla G et al. ENRICH trial. N Engl J Med. 2024;390(14):1277-89.

Remerciements : les annotateurs et l'équipe du laboratoire du Dr Létourneau-Guillon (CRCHUM), ainsi que les auteurs des outils publics évalués.
Financement :
IRSC, subvention Projet 2025–2029; bourse PREMIER, Faculté de médecine, Université de Montréal.
Laurent Létourneau-Guillon est soutenu par une bourse salariale Junior 2 du Fonds de recherche du Québec – Santé (FRQS)/Fondation de l'Association des radiologistes du Québec (FARQ) (https://doi.org/10.69777/380236).

Canadian Institutes of Health Research | Instituts de recherche en santé du Canada

Fonds de recherche Québec