Shared nnU-Net Probability Maps for msTBI Lesion Detection and Segmentation

Nair Ul Islam

GL Bajaj Institute of Technology and Management
nairulislam6@gmail.com

Abstract. Shared 2D nnU-Net lesion-probability maps can serve scan-level lesion detection and voxel-level lesion segmentation in moderate-to-severe traumatic brain injury, provided that each task uses its own decision rule and endpoint. A 2D nnU-Net trained on 572 T1-weighted MRI scans produced the shared maps. An 82-scan development cohort was used for checkpoint, model, and threshold selection and for cross-validated downstream analyses. On these development maps, a normalized predicted lesion-volume rule and a maximum-probability rule were at least as competitive as a logistic map-feature classifier for detection. In a retrospective paired analysis of 44 development cases, adaptive segmentation increased mean Dice from 0.5166 to 0.5541, a difference of +0.0375 (95% CI 0.0152–0.0611). On the hidden challenge set, the complete submitted detector achieved balanced accuracy 0.8762, with sensitivity 0.8313 and specificity 0.9211. The two blind segmentation aggregates were reported separately and were not paired; Dice was 0.4525 for standard segmentation and 0.1671 for adaptive segmentation. Shared probability maps therefore supported both tasks, but development analyses did not demonstrate an advantage for multifeature detection, and adaptive segmentation did not show external benefit in its separate blind evaluation.

Keywords: Traumatic brain injury · Lesion detection · Lesion segmentation · nnU-Net

1 Introduction

Moderate to severe traumatic brain injury is still a major clinical problem in adults [6]. Consequently, multi-site structural neuroimaging programs must cope with acquisition and population heterogeneity when they target reproducible imaging endpoints [8]. Under that pressure, T1-weighted decisions about lesion presence and lesion extent are consequential. The same scan may need a scan-level answer and a voxel-level answer, each with its own clinical and metric meaning.

AIMS-TBI places msTBI lesion analysis on T1-weighted MRI into a public dual-task setting with explicit lesion labels [2]. Detection asks whether a scan contains a lesion. Segmentation asks which voxels belong to it and how far the lesion extends. The two questions use different endpoints and decision rules, so a shared soft map is useful only when task-specific inference keeps those roles separate rather than collapsing them into one interchangeable score.

That separation motivates a shared representation with distinct decisions. We study a dual-task system built on one 2D nnU-Net as a self-configuring baseline backbone [3]. Here, a 2D nnU-Net trained once produces T1 lesion-probability maps that feed task-specific detection and segmentation rules for separate challenge submissions. Development analyses on the frozen maps explain how scan-level and voxel-level rules behave. Blind challenge evaluation supplies the external aggregates against which those explanations must be judged.

2 Related Work

Comparing task-specific inference from fixed maps requires separating upstream changes to the representation from downstream decision rules. Prior pipelines improve voxel maps, for example through adaptive intensity normalization restricted to brain parenchyma [9] or large-scale multi-dataset pretraining before fine-tuning on msTBI T1 scans [10]. Multi-sequence lesion CNNs such as DeepMedic are strong when several MRI contrasts are available [5]. With the trained network and its soft maps held fixed, we vary only the inference procedure per task, so the reported metrics attach to those inference choices rather than to a second training recipe, consistent with guidance to separate a proposed component from other training or inference confounds [4].

Once the representation is fixed, evaluation should follow the purpose of each decision. Scan-level detection is an image-level classification problem, whereas segmentation is a pixel-level problem, so metric choice should be task-specific [7]. Metric suitability and generalization across heterogeneous data conditions are related but different concerns. The Medical Segmentation Decathlon evaluates algorithm generalization across heterogeneous biomedical segmentation tasks [1]. In our study, this distinction motivates separate evidence roles for method development and blind evaluation.

3 Methods

The labeled pool contained 654 scans, of which 572 trained the segmentation backbone and 82 formed the development cohort used for specification, threshold, and rule selection. A scan was detection-positive when its total reference lesion volume exceeded 10 voxels, equivalently when its reference lesion mask contained more than 10 lesion voxels. The 82 development scans included two pairs of separately stored, byte-for-byte identical T1 images. Records in each pair stayed together whenever relevant cross-validation folds were assigned. For detection analysis and resampling, keeping each identical-image pair together gave 80 groups across the 82 scans; every remaining scan formed its own group. The original 572/82 training-development split did not place identical T1 images on opposite sides. This grouping controls identical images but provides no information about patient identity. Checkpoint selection for the shared representation

accessed the development cohort, so later internal analyses on frozen maps estimate performance conditional on that selected representation rather than as an untouched end-to-end test of the 82 scans.

Training used a single train/validation partition for 1000 epochs; the selected checkpoint supplied frozen soft lesion-probability maps for all downstream work. Detection used non-test-time-augmentation (non-TTA) maps. The blind standard-segmentation submission used non-TTA inference. After map export, analyses reused existing soft maps and hard masks without further network training or image inference.

The submitted detection system summarized each soft map with fourteen aggregate features and applied a logistic map-feature decision procedure. The original feature set comprised maximum probability; top-k mean probabilities for $k \in \{64, 256, 1024\}$; soft volumes at thresholds 0.10, 0.30, and 0.50; the 95th percentile of map probabilities; connected-component count, largest-component size, and component mass at threshold 0.30; and three transforms of those summaries (log soft volume, log largest-component size, and a top-k ratio). Features were robust-scaled. Class-balanced logistic regression with regularization strength $C = 0.05$ produced continuous scores that were thresholded at approximately 0.57695 (the median of resampling-partition development thresholds; equivalently ≈ 0.577).

Two further development analyses separated fixed specification from nested selection on the same frozen maps. Cross-validated development analysis with the classifier specification held fixed kept the fourteen-feature set and $C = 0.05$ constant. Before each held-out fold was scored, the logistic model was fitted and its balanced-accuracy threshold was selected using only the remaining development cases. Records from each identical T1-image pair stayed in the same fold. In nested development analysis, the remaining development cases were used to choose between the core and fourteen-feature sets and $C \in \{0.05, 0.2, 1\}$, fit the selected model, and select its threshold before the held-out fold was scored. Nesting therefore operated at the aggregation layer on frozen maps rather than as end-to-end network validation. Five-fold cross-validation was repeated three times. We report the first run in detail and summarize all three by their mean and standard deviation. These summaries were unpooled and describe sensitivity to the cross-validation splits.

Protocol-matched simple comparators on the same frozen maps and held-out cross-validated development predictions were a maximum-probability rule and a normalized predicted lesion-volume rule. The maximum-probability rule used the maximum voxel probability on each map as a continuous score. The normalized predicted lesion-volume rule used the fraction of voxels with probability strictly greater than 0.5, that is, the number of such voxels divided by the total number of voxels in the map. For both simple rules, operating thresholds for binary decisions were selected by maximizing balanced accuracy on predictions for the remaining development cases before the held-out fold was evaluated, using the same threshold-selection procedure as for the logistic analyses.

In the retrospective paired analysis, standard segmentation thresholded the TTA maps at $p \geq 0.5$, and adaptive segmentation operated on the same maps. The adaptive rule first counted voxels with probability $p \geq 0.5$. Counts at or below 10,000 entered a low-burden branch that thresholded at $p \geq 0.002$ and pruned every 26-connected predicted component of at most 10 voxels. Larger counts used the $p \geq 0.5$ mask. The submitted thresholds and pruning values were development-stage choices rather than parameters derived from an independent tuning cohort or a prespecified optimization procedure. We therefore evaluate their sensitivity retrospectively rather than treating the submitted values as independently optimized. Retrospective paired segmentation analysis applied both policies to the same 44 cases, TTA maps, and filtered references. For challenge segmentation evaluation, those references retained only 26-connected reference-mask components strictly larger than 50 voxels; this eligibility rule was not part of adaptive prediction. Of the 82 development scans, 35 were detection-negative and three detection-positive scans lacked an eligible reference component, leaving 44 cases under the same filter for both methods. The 44 cases formed 43 groups because one identical-T1 pair remained together; every other case formed its own group. A nested adaptive sensitivity analysis explored a retrospective 160-rule grid (eight low thresholds $\times$ five component minima $\times$ four routing cutoffs), selecting a rule by mean Dice on the remaining development cases before scoring the held-out fold. That grid is a sensitivity design, not confirmatory optimization of an external adaptive claim.

Detection metrics were area under the ROC curve, balanced accuracy, sensitivity, and specificity. Segmentation used Dice as the primary overlap measure, with surface distances available when reported later. The Dice denominator in the original analysis comprised 60 cases with defined scores, namely all 47 lesion-containing scans plus 13 detection-negative scans that produced false-positive masks. The remaining 22 empty-reference and empty-prediction pairs had undefined Dice and were excluded from that finite mean. Probability diagnostics for development scores included Brier score, expected calibration error (ECE), and calibration slope, with nested Platt scaling as a sensitivity check. Brier score was the mean squared error between labels and predicted probabilities. ECE was computed with eight equal-count bins obtained by sorting predicted probabilities and splitting the ordered indices into eight groups of as equal size as possible, then taking the size-weighted mean absolute difference between bin-wise mean predicted probability and bin-wise observed positive rate. Calibration slope was obtained from a logistic regression of the label on the logit of the predicted probability. Where intervals apply, detection inference used 10,000 bootstrap resamples of all 80 groups, keeping each identical-image pair together and stratifying by class for contrasts when appropriate. Sensitivity and specificity used Clopper–Pearson exact intervals. Paired segmentation deltas used grouped bootstrap and grouped sign permutation across 43 groups, keeping the one remaining identical-image pair together; every other case remained an individual group. Lesion-size strata were descriptive only. Blind challenge evaluation reports only the aggregates supplied separately for detection, non-TTA standard segmentation, and adaptive

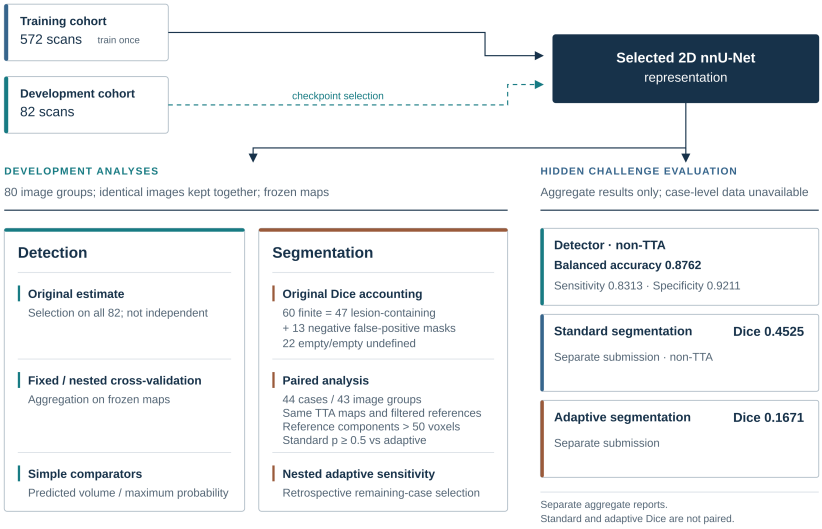

Fig. 1. Evaluation topology. Development analyses and blind challenge evaluations form parallel branches from the selected representation. In the paired development comparison, the standard threshold-0.5 rule and the adaptive rule use identical TTA maps, cases, and references.

segmentation. Hidden labels, case-level predictions, and challenge denominators were unavailable for further blind re-analysis.

4 Results

On the hidden challenge set, the submitted detection system achieved balanced accuracy 0.8762, with sensitivity 0.8313 and specificity 0.9211. These are the aggregate values reported for the submitted detector on the hidden challenge set; the challenge's official final ranking uses a separate procedure. Case-level predictions, challenge denominators, and a blind AUC were not supplied, so the report remains aggregate-only.

The original development estimate, obtained after specification and threshold selection on the 82-case cohort, was AUC 0.9392.

Cross-validated development analysis with the classifier specification held fixed produced AUC 0.9277 (95% CI 0.8634–0.9757), balanced accuracy 0.8900, sensitivity 0.8085 (0.6674–0.9085), and specificity 0.9714 (0.8508–0.9993). Across three repeats, mean AUC was 0.9321 (SD 0.0039). Nested development analysis, with aggregation-layer selection using the remaining development cases before each held-out cross-validation fold was scored on the same frozen maps, produced AUC 0.9252 (0.8604–0.9740) and balanced accuracy 0.9043; the three-repeat mean AUC was 0.9307 (SD 0.0052).

Table 1. Detection results by evaluation role.

Evaluation	Cases/data role	Method	AUC (95% CI)	Balanced accuracy	Sensitivity	Specificity
Blind challenge evaluation	Hidden challenge; aggregate report only	Submitted detection system	–	0.8762	0.8313	0.9211
Original development estimate	82-case development; after specification and threshold selection	Logistic map-feature classifier	0.9392	–	–	–
Cross-validated development analysis, fixed classifier specification	82 cross-validated development predictions; 80 image groups	Logistic map-feature classifier	0.9277 (0.8634–0.9757)	0.8900	0.8085 (0.6674–0.9085)	0.9714 (0.8508–0.9993)
Nested development analysis	82 cross-validated development predictions; selection on remaining development cases using frozen maps	Logistic map-feature classifier	0.9252 (0.8604–0.9740)	0.9043	–	–
Normalized predicted lesion volume	Same cross-validated development predictions	Normalized predicted lesion-volume rule	0.9416	0.9043	–	1.000 (0.9000–1.0000)
Maximum-probability comparator	Same cross-validated development predictions	Maximum-probability rule	0.9398	0.8827	–	–

Protocol-matched simple rules on the same held-out cross-validated development maps remained competitive. The normalized predicted lesion-volume rule achieved AUC 0.9416, balanced accuracy 0.9043, and observed specificity 1.000 (exact CI 0.9000–1.0000). The maximum-probability rule achieved AUC 0.9398 and balanced accuracy 0.8827. The fixed logistic map-feature classifier, for comparison, had AUC 0.9277 and balanced accuracy 0.8900. The normalized predicted lesion-volume rule AUC was +0.0140 higher than the logistic map-feature classifier AUC (paired 95% CI −0.0006 to 0.0377). Because that interval crosses zero, the internal analyses did not demonstrate a multifeature advantage. Supplementary Tables S1 and S2 report the remaining comparators and ablations, their uncertainty intervals, and multiplicity-adjusted comparisons.

Cross-validated development scores from the fixed logistic analysis had Brier 0.1143, ECE 0.1274, and slope 0.509. Scores from the nested Platt sensitivity analysis had Brier 0.1155, ECE 0.1319, and slope 0.457. Nested Platt scaling did not improve the reported diagnostics. Supplementary Table S3 gives log loss and calibration intercepts, and Supplementary Figure S1 plots the eight equal-count reliability bins.

The original development summary on the 82-case export gave lesion-containing mean Dice 0.517638 on the 47 positive scans and finite-case mean

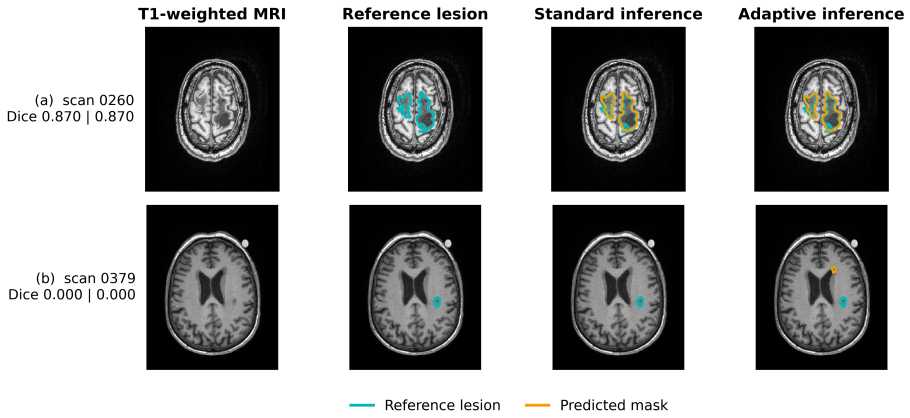

Fig. 2. Highest-Dice case and a complete miss under standard and adaptive segmentation. Cyan shows the reference lesion; orange shows the predicted mask.

Dice 0.405483 on 60 defined values. The finite set comprises the 47 positives plus 13 detection-negative scans with false-positive masks.

Retrospective paired segmentation analysis used 44 cases that formed 43 groups after the identical-image pair was kept together under a shared reference filter. The standard $p \geq 0.5$ rule and the adaptive rule were applied to the same TTA maps and filtered references. Standard segmentation Dice was 0.5166 and adaptive segmentation Dice was 0.5541, a paired difference of +0.0375 (95% CI 0.0152–0.0611; grouped permutation $p = 0.0027$). Empty-prediction misses on lesion-containing cases moved from 1 under standard to 0 under adaptive. HD95 and ASSD differences were not statistically resolved, so no surface-distance improvement is claimed. Descriptive lesion-size strata showed Dice deltas of +0.0493 for 51–500 voxels ($n = 9$), +0.0696 for 501–2000 ($n = 13$), and +0.0136 for volumes above 2000 voxels ($n = 22$). These strata are descriptive only; no interaction test is reported. The examples in Fig. 2 show the highest-Dice case under both rules and one case missed by both.

A nested adaptive sensitivity analysis, selecting among a retrospective 160-rule grid using the remaining development cases before each held-out cross-validation fold was scored, produced mean Dice 0.5438 versus 0.5166 for standard segmentation on the same 44 cases (delta +0.0272, 95% CI 0.0049–0.0509, $p = 0.0298$). That comparison is a sensitivity check, not confirmatory optimization of an external adaptive claim.

Under blind challenge evaluation, the separate non-TTA standard submission received aggregate Dice 0.4525, and the adaptive submission received aggregate Dice 0.1671. Those values come from separate aggregate reports and are not a case-level paired external comparison.

Table 2. Segmentation results by evaluation role and denominator.

Evaluation	Cases/data role	Method	Dice
Original development summary	47 lesion-containing scans	Standard segmentation, original summary	0.517638
Original development summary	60 finite values (47 positives + 13 negative false-positive masks); 22 empty/empty undefined	Standard segmentation, original summary	0.405483
Retrospective paired segmentation analysis	44 cases / 43 image groups	Standard segmentation, $p \geq 0.5$ on TTA maps	0.5166
Retrospective paired segmentation analysis	44 cases / 43 image groups	Adaptive segmentation on the same TTA maps	0.5541
Blind challenge evaluation	Hidden challenge; aggregate report only	Standard segmentation, separate non-TTA submission	0.4525
Blind challenge evaluation	Hidden challenge; aggregate report only	Adaptive segmentation, separate submission	0.1671

5 Discussion

What the development analyses taught is a bound on attribution. The normalized predicted lesion-volume rule AUC was +0.0140 higher than the logistic map-feature classifier AUC, but its confidence interval crossed zero, so no multifeature advantage was demonstrated and neither superiority nor equivalence follows. The blind challenge score therefore cannot be attributed specifically to the 14-feature logistic aggregation layer.

What happened with segmentation depends on the evaluation setting. In the retrospective paired analysis, adaptive inference raised Dice by +0.0375 on matched development cases, a development-only gain under rules chosen on that cohort. Blind aggregates, reported separately, gave standard Dice 0.4525 and adaptive Dice 0.1671. Because those blind numbers are not a case-level paired comparison, they cannot confirm an external adaptive benefit; the adaptive blind result instead signals development-specific tuning risk. Multi-fold cross-validation would be preferable for more reliable segmentation benchmarking [4].

6 Limitations

Several bounds limit how far the present results can be generalized. Only one 2D nnU-Net representation was trained, and variability from network retraining was not measured. Post-review repeated cross-validation therefore assesses the aggregation layer on frozen maps rather than nnU-Net retraining. Checkpoint

selection also accessed the 82-case development cohort. Accordingly, these internal estimates remain conditional on the selected representation and should not be interpreted as performance from a fully untouched development cohort.

The available metadata did not include a verified patient identifier, so patient-level independence could not be assessed. Keeping identical images together controls duplicate-image placement without establishing patient identity. The same metadata did not include verified site labels, so no site-stratified evaluation was performed or reported. Only aggregate challenge results were available. Case-level predictions were unavailable; therefore, subgroup analyses, calibration assessment on hidden data, AUC reconstruction, and a paired external standard-versus-adaptive comparison could not be performed.

7 Conclusion

Shared 2D nnU-Net lesion-probability maps support both scan-level detection and voxel-level segmentation, but the strength of evidence differs by task and evaluation setting. The complete submitted detector was assessed on the hidden challenge set. Development analyses did not demonstrate a multifeature advantage over normalized predicted lesion volume and maximum-probability rules. Adaptive segmentation improved matched development Dice yet showed no external blind benefit (Dice 0.1671). Future work using a verified patient identifier to assess patient-level independence, together with repeated representation training, would tighten those bounds.

8 Code and data availability

Code and analysis scripts will be made available at `https://github.com/NairUlIslam/shared-nnunet-probability-maps-msTBI`.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., Bilello, M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M.J., Heckers, S.H., Huisman, H., Jarnagin, W.R., McHugo, M.K., Napel, S., Pernicka, J.S.G., Rhode, K., Tobon-Gomez, C., Vorontsov, E., Meakin, J.A., Ourselin, S., Wiesenfarth, M., Arbeláez, P., Bae, B., Chen, S., Daza, L., Feng, J., He, B., Isensee, F., Ji, Y., Jia, F., Kim, I., Maier-Hein, K., Merhof, D., Pai, A., Park, B., Perslev, M., Rezaiifar, R., Rippel, O., Sarasua, I., Shen, W., Son, J., Wachinger, C., Wang, L., Wang, Y., Xia, Y., Xu, D., Xu, Z., Zheng, Y., Simpson, A.L., Maier-Hein, L., Cardoso, M.J.: The Medical Segmentation Decathlon. Nature Communications **13**(1), 4128 (Jul 2022). `https://doi.org/10.1038/s41467-022-30695-9`, `https://www.nature.com/articles/s41467-022-30695-9`

2. Dennis, E., Deutscher, E., Onicas, A., Bakas, S., Pease, M., Baid, U., Wilde, E.: AIMS-TBI : Automated Identification of Moderate-Severe Traumatic Brain Injury Lesions (Apr 2026). https://doi.org/10.5281/zenodo.19731996, https://zenodo.org/records/19731996
3. Isensee, F., Jäger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: Automated Design of Deep Learning Methods for Biomedical Image Segmentation. Nature Methods **18**(2), 203–211 (Feb 2021). https://doi.org/10.1038/s41592-020-01008-z, http://arxiv.org/abs/1904.08128, arXiv:1904.08128 [cs.CV]
4. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jäger, P.F.: nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 488–498. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72114-4_47
5. Kamnitsas, K., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis **36**, 61–78 (Feb 2017). https://doi.org/10.1016/j.media.2016.10.004, https://www.sciencedirect.com/science/article/pii/S1361841516301839
6. Maas, A.I.R., Stocchetti, N., Bullock, R.: Moderate and severe traumatic brain injury in adults. The Lancet. Neurology **7**(8), 728–741 (Aug 2008). https://doi.org/10.1016/S1474-4422(08)70164-9
7. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M.A., Wiesenfarth, M., Kavur, A.E., Sudre, C.H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M.B., Cardoso, M.J., Cheplygina, V., Cimini, B.A., Collins, G.S., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Haase, R., Hashimoto, D.A., Hoffman, M.M., Huisman, M., Jannin, P., Kahn, C.E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B.A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A.L., Mattson, P., Meijering, E., Menze, B., Moons, K.G.M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C.I., Shetty, S., van Smeden, M., Summers, R.M., Taha, A.A., Tiulpin, A., Tsaftaris, S.A., Van Calster, B., Varoquaux, G., Jäger, P.F.: Metrics reloaded: recommendations for image analysis validation. Nature Methods **21**(2), 195–212 (Feb 2024). https://doi.org/10.1038/s41592-023-02151-z, https://www.nature.com/articles/s41592-023-02151-z
8. Olsen, A., Babikian, T., Bigler, E.D., Caeyenberghs, K., Conde, V., Dams-O'Connor, K., Dobryakova, E., Genova, H., Grafman, J., Håberg, A.K., Heggland, I., Hellstrøm, T., Hodges, C.B., Irimia, A., Jha, R.M., Johnson, P.K., Koliatsos, V.E., Levin, H., Li, L.M., Lindsey, H.M., Livny, A., Løvstad, M., Medaglia, J., Menon, D.K., Mondello, S., Monti, M.M., Newcombe, V.F.J., Petroni, A., Ponsford, J., Sharp, D., Spitz, G., Westlye, L.T., Thompson, P.M., Dennis, E.L., Tate, D.F., Wilde, E.A., Hillary, F.G.: Toward a global and reproducible science for brain imaging in neurotrauma: the ENIGMA adult moderate/severe traumatic brain injury working group. Brain Imaging and Behavior **15**(2), 526–554 (Apr 2021). https://doi.org/10.1007/s11682-020-00313-7
9. Son, I., Lee, G., Sim, S., Uhm, K.H.: Lesion Segmentation in Moderate to Severe Traumatic Brain Injury: An NnU-Net Based Approach with Adaptive Normalization

in the AIMS-TBI 2025 Challenge. In: Bakas, S., Dennis, E., Astaraki, M., Baid, U., Conte, G.M., Foltyn-Dumitru, M., Jiang, Z., Labella, D., Metz, M.C., Anazodo, U., de Verdier, M.C., Kofler, F., Li, H.B., Linguraru, M.G., Maleki, N. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. pp. 322–329. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-16370-7_29

10. Ulrich, C., Wald, T., Isensee, F., Maier-Hein, K.H.: Large Scale Supervised Pretraining For Traumatic Brain Injury Segmentation (Apr 2025). https://doi.org/10.48550/arXiv.2504.06741, http://arxiv.org/abs/2504.06741, arXiv:2504.06741 [cs.CV]