Conformal 3D Lesion Segmentation with Balanced Risk Control

Binyu Tan[1⋆], Zhiyuan Wang[1⋆], Jinhao Duan[2], Kaidi Xu[3], Heng Tao Shen[4], Fumin Shen[1], and Xiaoshuang Shi[1(✉)]

[1] University of Electronic Science and Technology of China, Chengdu, China
xsshi2013@gmail.com
[2] University of North Carolina at Chapel Hill, Chapel Hill, USA
[3] City University of Hong Kong, Hong Kong, China
[4] Tongji University, Shanghai, China

Abstract. Medical image segmentation models commonly binarize voxel-wise scores using a fixed threshold, such as 0.5. However, this heuristic provides no finite-sample guarantee against clinically consequential under-segmentation, particularly in 3D lesion segmentation, where lesions are sparse relative to the background. We propose Conformal Lesion Segmentation (CLS), a model-agnostic post-processing framework for controlling voxel-level false-negative risk. Given a user-specified FNR tolerance ε and violation probability α, CLS uses a held-out calibration set to compute, for each case, the smallest mask-expansion parameter for which the voxel-level FNR does not exceed ε. A finite-sample-corrected conformal quantile of these critical values is then used to determine a global binarization threshold. Under exchangeability, CLS guarantees that a new test case satisfies the prescribed FNR tolerance with probability at least $1 - \alpha$. Because the critical parameter is chosen as the smallest feasible expansion, CLS also limits the associated increase in false positives, although FPR is not itself formally controlled. We evaluate CLS on six 3D-LS benchmarks across five backbone models, demonstrating its superior statistical validity and predictive performance, and providing potential guidance for deploying risk-aware segmentation models in real-world clinical applications. Code can be found at https://github.com/binyutan/CLS.

Keywords: 3D lesion segmentation · Conformalization.

1 Introduction

Recent advances in deep learning and computer vision [7] have enabled numerous automated medical image segmentation models for modalities such as CT and MRI [18,22], achieving expert-level performance across diverse clinical tasks [27]. Despite these gains, current models typically depend on a fixed, heuristic threshold (i.e., 0.5) to delineate target structures from background,

⋆ Co-first authors.

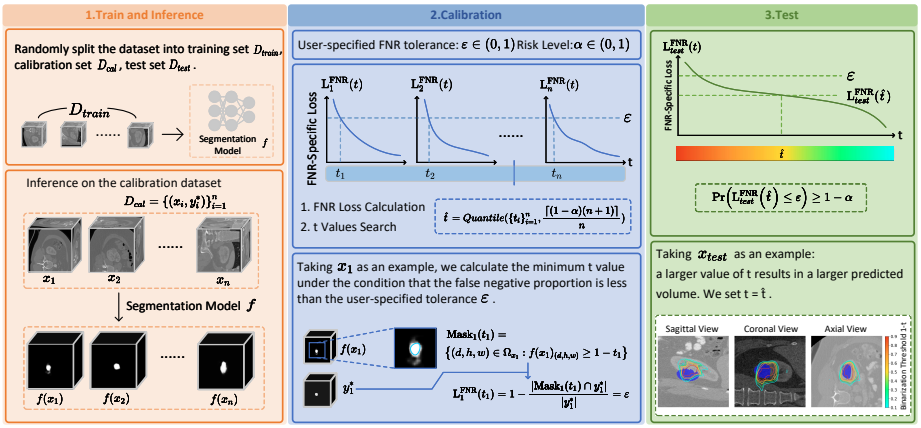


Fig. 1: Overview of the CLS framework.

lacking statistically valid guarantees for critical safety metrics such as the false negative rate (FNR) [12]. This limitation compromises their trustworthiness in risk-sensitive clinical environments, where robust risk control is essential [18]. In particular, under-segmentation can cause the model to either incompletely delineate lesion boundaries or entirely miss lesions, resulting in pathological regions being missed and thus left untreated [15,29]. For instance, in early-stage tumor screening, missing lesions smaller than 5 mm can have serious consequences for patient outcomes, as false negatives directly compromise diagnostic and therapeutic decisions [16,17]. These concerns highlight the need for statistically grounded segmentation frameworks, especially in challenging scenarios such as 3D lesion segmentation (3D-LS) [19,9].

Split (inductive) conformal prediction (SCP) [20,5,25] has recently emerged as a principled solution to address these shortcomings. SCP offers distribution-free, model-agnostic guarantees on ground-truth coverage, assuming data exchangeability. This framework reserves a held-out calibration set to compute nonconformity scores that quantify the discrepancy between model predictions and ground-truth labels. By computing the quantile of these scores at a user-specified risk level α as the test-time selection threshold, SCP then constructs a prediction set for each test sample ensuring that the true label is included with probability at least $1-\alpha$. While SCP has demonstrated strong performance in classification settings [1], its adaptation to segmentation for controlling per-case voxel-level FNR has received limited attention.

In this paper, we introduce Conformal Lesion Segmentation (CLS), a novel extension of SCP to image segmentation, which constrains the test-time voxel-level FNR (V-FNR) to lie below a user-specified tolerance ε with high probability $1-\alpha$, while maintaining a low V-FPR. As shown in Figure 1, CLS builds upon a pretrained segmentation model and focuses on designing a task-specific nonconformity score that reflects lesion-level false negative risk under the given FNR

constraint. Specifically, for each calibration sample, CLS identifies the maximum binarization threshold $1-t$ such that the V-FNR remains below the target tolerance ε. The collection of these per-sample critical thresholds over the calibration set forms a distribution of nonconformity scores, from which CLS computes the approximate $1-\alpha$ quantile to determine a statistically calibrated test-time threshold. This calibrated threshold ensures, with probability at least $1-\alpha$, that the FNR on unseen test examples remains within the user-defined tolerance ε.

We evaluate CLS on six 3D-LS benchmarks utilizing five popular 3D medical image segmentation models. Empirical results demonstrate that CLS consistently enforces the FNR constraint within the predefined tolerance ε at the user-specified risk level α. By rigorously calibrating the test-time threshold, CLS achieves significantly lower FNR on both presence-level and voxel-level compared to those obtained employing a heuristic threshold of 0.5 across all datasets. Beyond FNR control, we further examine how predicted region sizes vary across models and risk levels, offering a practical and interpretable tool for benchmarking uncertainty-aware segmentation models.

Our main contributions are summarized as follows:

- We propose Conformal Lesion Segmentation (CLS) that extends SCP to binary segmentation settings, addressing a clinically driven risk objective: bounding the probability that per-case voxel-level FNR exceeds a user-specified tolerance via a single calibrated post-processing threshold, while preserving standard binary mask output.
- We derive novel nonconformity measures from false negative risk-constrained critical thresholds, facilitating statistically rigorous segmentation with FNR control and inherently low FPR.
- We introduce *prediction compactness* as an auxiliary metric that complements V-FNR, P-FNR, and V-FPR, summarizing the spatial footprint required to achieve a given FNR tolerance and enabling interpretable comparison of backbone conservativeness under risk-controlled calibration.

2 Method

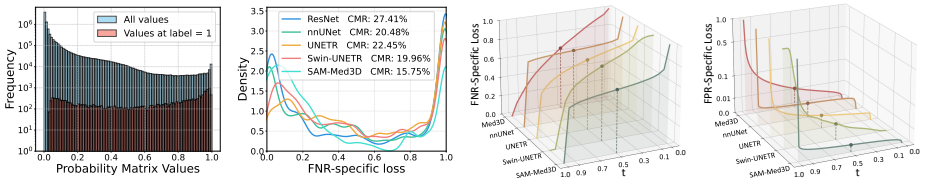
2.1 Notations and Problem Formulation

Formally, the dataset is partitioned into three disjoint subsets: a training set $\mathcal{D}_{train}$, a calibration set $\mathcal{D}_{cal}$, and a test set $\mathcal{D}_{test}$. Since SCP is model-agnostic, we first fit a segmentation model $f(\cdot)$ on $\mathcal{D}_{train}$. Given a 3D image $x_i \in \mathbb{R}^{D\times H\times W}$, the model outputs a confidence map $\hat{y}_i = f(x_i) \in [0,1]^{D\times H\times W}$, where each voxel value denotes the predicted probability of lesion presence. The corresponding ground-truth annotation is a binary mask $y_i^* \in \{0,1\}^{D\times H\times W}$.

For a decision threshold $t \in [0,1]$, the predicted lesion region is defined as

$$\mathrm{Mask}_i(t) = \left\{(d,h,w) \in \Omega_{x_i} : f(x_i)_{(d,h,w)} \geq 1-t\right\}, \tag{1}$$

where $\Omega_{x_i} \subset \mathbb{Z}^3$ denotes the spatial index set of x_i. Voxels with confidence scores above $1-t$ are classified as lesion, and the remainder as background.



(a) Prob. Value Dist. (b) V-FNR Dist. (c) FNR Loss vs. t (d) FPR Loss vs. t

Fig. 2: Distributions and Voxel-level FNR/FPR Loss vs. t

As discussed above, heuristically adopting a fixed decision threshold $1-t$ (e.g., 0.5) provides no formal control over false negative risk. Figure 2a illustrates the average distribution of output probabilities produced by the Med3D model [8] on the KiTS21 dataset [13], where a large fraction of lesion voxels (label $= 1$) receive predicted probabilities below 0.5. Figure 2b further shows the distribution of voxel-level FNR (V-FNR) across models, with presence-level FNR (P-FNR) annotated; notably, many samples exhibit V-FNR $= 1$, indicating complete lesion miss. Consistently, as shown in Figure 2c-2d, at the conventional threshold $t = 0.5$, the average V-FNR (0.6023) is nearly two orders of magnitude larger than the average V-FPR (0.0082), revealing a strong bias toward false negatives. These results demonstrate that a fixed threshold is inadequate across samples.

Our objective is therefore to learn a statistically valid threshold $1 - \hat{t}$ from a calibration set $\mathcal{D}_{cal} = \{(x_i, y_i^*)\}_{i=1}^n$ such that the test-time false negative risk is bounded by a user-specified tolerance ε with high probability, while keeping false positives low:

$$\Pr\left(R(\hat{t}) \leq \varepsilon\right) \geq 1 - \alpha, \tag{2}$$

where $R(\hat{t})$ denotes the FNR on unseen test samples, and α controls the allowable violation probability.

2.2 Conformal Lesion Segmentation

Exchangeable data distribution. As a foundational yet non-restrictive assumption, we posit that the n calibration samples $\{(x_i, y_i^*)\}_{i=1}^n$ and each test point (x_{test}, y_{test}^*) in $\mathcal{D}_{test}$ are exchangeable, which underlies the theoretical validity of SCP-based approaches [2].

To align the nonconformity score with the underlying false negative risk, we define a *threshold-dependent voxel-level FNR-specific loss function* for each calibration data point:

$$\mathrm{L}_i^{\mathrm{FNR}}(t) = 1 - \frac{|\mathrm{Mask}_i(t) \cap y_i^*|}{|y_i^*|} = 1 - \frac{\left|\left\{\substack{(d,h,w)\in\Omega_{x_i}:\\ f(x_i)_{(d,h,w)}\geq 1-t,\ y^*_{i(d,h,w)}=1}\right\}\right|}{\left|\left\{(d,h,w)\in\Omega_{x_i} : y^*_{i(d,h,w)} = 1\right\}\right|}, \tag{3}$$

where $y^*_{i(d,h,w)}$ is the ground-truth label at voxel (d, h, w) and a lower loss indicates that a larger fraction of the ground-truth lesion voxels are successfully

identified by the model. Analogously, we define a *threshold-dependent voxel-level FPR-specific loss function*:

$$\mathrm{L}_i^{\mathrm{FPR}}(t) = \frac{|\mathrm{Mask}_i(t) \cap (1 - y_i^*)|}{|1 - y_i^*|} = \frac{\left|\left\{ \substack{(d,h,w)\in\Omega_{x_i}: \\ f(x_i)_{(d,h,w)}\geq 1-t,\ y^*_{i(d,h,w)}=0} \right\}\right|}{\left|\left\{(d,h,w) \in \Omega_{x_i} : y^*_{i(d,h,w)} = 0\right\}\right|}, \tag{4}$$

A lower loss indicates a lower proportion of ground-truth background voxels being erroneously identified as lesions.

The monotonicity of the FNR-specific loss with respect to t is immediate from its definition, as also illustrated in Figure 2c. Leveraging this property, we then develop the *nonconformity score* as

$$t_i = \arg \min_{t:\mathrm{L}_i^{\mathrm{FNR}}(t)\leq\varepsilon} \mathrm{L}_i^{\mathrm{FPR}}(t) = \inf\left\{t : \forall t' \geq t, \mathrm{L}_i^{\mathrm{FNR}}(t') \leq \varepsilon\right\}, \tag{5}$$

which represents the lowest feasible decision threshold for each calibration sample under which the false negative proportion remains within the specified tolerance ε. This critical point t_i minimizes the predicted lesion area $\mathrm{Mask}_i(t_i)$ to minimize the false positive rate and enhance precision subject to the risk constraint. Subsequently, we sort all nonconformity scores $\{t_i\}_{i=1}^n$ on the calibration set in ascending order such that $t_1 \leq \cdots \leq t_n$, and compute their $\frac{\lceil (1-\alpha)(1+n) \rceil}{n}$ quantile:

$$\hat{t} = \inf\left\{t : \frac{|\{i : t_i \leq t\}|}{n} \geq \frac{\lceil (1-\alpha)(1+n) \rceil}{n}\right\}. \tag{6}$$

Theorem 1 (Statistically rigorous FNR constraint)**.** *Suppose the given test instance* (x_{test}, y^*_{test}) *and* $(x_i, y_i^*)_{i=1,\cdots,n}$ *are exchangeable, we employ* $\hat{t}$ *as the test-time decision threshold and the resulting predicted lesion region is* $\mathrm{Mask}_{test}(\hat{t})$*. Then the false negative proportion* $\mathrm{L}_{test}^{FNR}\left(\hat{t}\right)$ *satisfies*

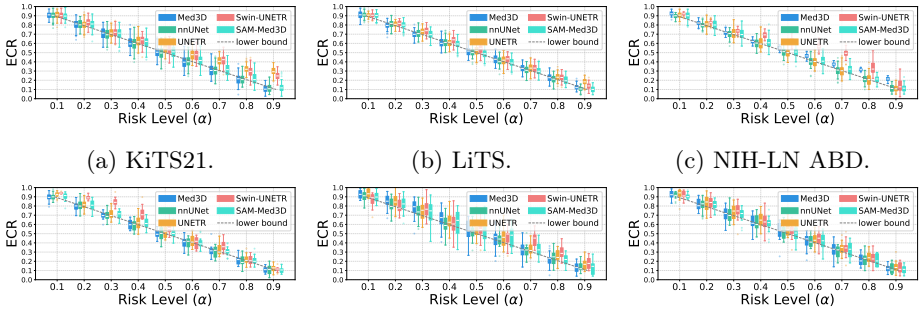
$$\Pr\left(\mathrm{L}_{test}^{\mathrm{FNR}}\left(\hat{t}\right) \leq \varepsilon\right) \geq 1 - \alpha. \tag{7}$$

This is the same property of risk control as Eq. (2). We emphasize that this is a *marginal*, per-test-sample guarantee: for any single exchangeable test instance, the probability of FNR exceeding ε is at most α. It does not imply that the empirical compliance rate (ECR) over every finite test partition must exceed $1 - \alpha$. Below, we establish its statistical rigor.

Proof of Theorem 1. Under the condition that $t_{i=1,\cdots,n}$ are in ascending order, $\hat{t}$ can be reformulated as $\hat{t} = t_{n\cdot\frac{\lceil (1-\alpha)(1+n)\rceil}{n}} = t_{\lceil (1-\alpha)(1+n)\rceil}$. By the definition of t_i, if $\mathrm{L}_{test}^{\mathrm{FNR}}\left(\hat{t}\right) \leq \varepsilon$, it can obtain $\hat{t} \geq t_{test}$. Since $(x_1, y_1^*), \cdots, (x_n, y_n^*), (x_{test}, y^*_{test})$ are supposed to be exchangeable, it has $\Pr\left(t_{test} \leq t_i\right) = \frac{i}{n+1}$. Finally, it can obtain

$$\Pr\left(\mathrm{L}_{test}^{\mathrm{FNR}}\left(\hat{t}\right) \leq \varepsilon\right) = \Pr\left(\hat{t} \geq t_{test}\right) = \frac{\lceil (1-\alpha)(1+n) \rceil}{n+1} \geq 1 - \alpha. \tag{8}$$

This completes the proof of Theorem 1 and establishes the statistical validity of the test-time decision threshold.



(a) KiTS21. (b) LiTS. (c) NIH-LN ABD.

(d) LIDC-IDRI. (e) MDSC-Colon. (f) MDSC-Pancreas.

Fig. 3: Test-time ECR results on six datasets utilizing five segmentation models.

3 Experiments

3.1 Experimental Settings

Datasets. We consider six fully annotated 3D medical segmentation datasets from the ULS23 Segmentation Challenge [9], each covering a different anatomical region or organ system: KiTS21 [13], LiTS [6], NIH-LN ABD [21], LIDC-IDRI [4], MDSC-Colon [3], and MDSC-Pancreas [3].

Backbone Models. We employ five representative 3D medical image segmentation models with diverse architectural designs, each fine-tuned to maintain a comparable parameter scale: Med3D[8], nnUNet[14], UNETR[11], Swin-UNETR[10], and SAM-Med3D[23].

Evaluation Metrics. We check whether the ***empirical compliance rate*** (ECR) [2], defined as the proportion of test samples whose FNR-specific loss is controlled below ε, exceeds $1-\alpha$. We use the ***presence-level FNR*** (P-FNR), which measures the proportion of samples where lesions are completely missed (voxel-level FNR $= 1$), alongside ***V-FNR*** and ***V-FPR*** for comprehensive assessment. To quantify spatial precision under risk control, we introduce ***prediction compactness*** (PC): the ratio of predicted lesion voxels to total image voxels. Under the FNR below ε constraint, true positive volume is guaranteed; thus, lower PC implies fewer false positives, yielding lower FPR and higher Dice. PC provides an efficient, interpretable metric for comparing spatial efficiency across models and risk levels α.

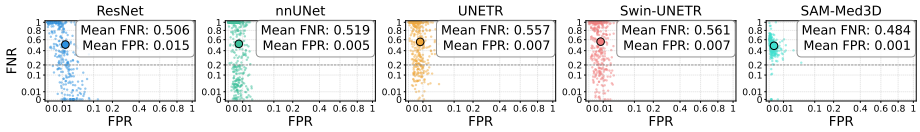
3.2 Empirical Evaluations

We fix the calibration-test split ratio at 0.5 by default. Each experimental group is evaluated over 100 random splits.

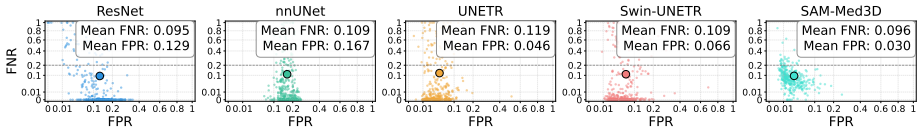
Statistical Validity of CLS. We first validate the statistical rigor of Theorem 1. As illustrated in Figure 3, thresholds calibrated via Eq. (6) consistently control test-time ECR under target levels across all six 3D-LS datasets and

Table 1: Test-time V-FNR and P-FNR on KITS21 using heuristic ($t = 0.5$) and CLS-calibrated thresholds at varying tolerance levels (ε).

Models	**Med3D**		**UNETR**		**SAM-Med3D**	
ε	**V-FNR**	**P-FNR**	**V-FNR**	**P-FNR**	**V-FNR**	**P-FNR**
$t = 0.5$ (Fixed by default)						
	0.5058	0.2741	0.5573	0.2545	0.4835	0.1575
$t = \hat{t}$ calibrated at the risk level of 0.2						
0.1	0.0533	0.0210	0.0576	0.0159	0.0469	0.0000
0.3	0.1654	0.0419	0.1693	0.0553	0.1779	0.0257
0.5	0.3211	0.0975	0.3528	0.0815	0.2995	0.0825



(a) Voxel-level FNR–FPR distribution using a heuristic threshold of 0.5.



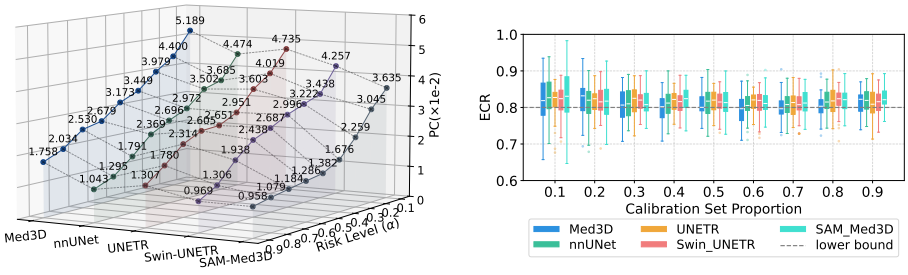
(b) Voxel-level FNR–FPR distribution after CLS calibration ($\varepsilon = 0.2$, $\alpha = 0.2$)

Fig. 4: Voxel-level FNR–FPR distribution before and after calibration.

five pretrained segmentation models. Notably, Theorem 1 provides a *marginal* guarantee for each individual test sample; ECR is an empirical average over a finite test partition and is not itself guaranteed to exceed $1 - \alpha$ on every split. Accordingly, occasional ECR values falling slightly below $1 - \alpha$ across the 100 random splits do not contradict the conformal guarantee but reflect the inherent variability of finite-sample empirical estimates [2,26,28,24].

Comparison with Heuristic Thresholding. As presented in Table 1, under a fixed threshold of 0.5, all three pretrained segmentation models exhibit substantially high test-time FNRs (both voxel-level and presence-level) on KiTS21. Notably, the P-FNR remains alarmingly high across models, indicating that a non-trivial proportion of lesions are completely missed, which poses substantial clinical risk in safety-critical settings. By contrast, CLS derives statistically rigorous thresholds under a user-specified risk level (e.g., $\alpha = 0.2$), yielding consistent reductions in FNR across all models and V-FNR tolerance levels ε.

Multi-metric Trade-off. Figure 4 illustrates the expected trade-off between false negatives and false positives. Under the heuristic threshold of 0.5, many cases lie in the high-FNR, low-FPR region, including complete lesion misses. CLS



(a) PC results on KiTS21. (b) ECR results using Med3D on KiTS21.

Fig. 5: PC results on KiTS21 and ECR results using Med3D on KiTS21 under various calibration-to-test splits, with $\varepsilon = 0.4$ and $\alpha = 0.2$.

shifts the distribution toward substantially lower FNR, at the cost of an increase in FPR caused by the enlarged predicted masks. Importantly, CLS selects the smallest threshold expansion required by its calibration criterion, thereby limiting, rather than eliminating, the accompanying increase in false positives. These results show that CLS prioritizes clinically important false-negative control while retaining an interpretable FNR–FPR trade-off.

Prediction Compactness as a Risk-aware Benchmark. To characterize performance beyond FNR control, we evaluate PC on the KiTS21 dataset across five segmentation models under varying risk levels. As shown in Figure 5a, higher α leads to less conservative and more spatially compact predictions. Moreover, flatter curves indicate that the calibrated mask size is relatively insensitive to the chosen risk level, whereas steeper curves indicate stronger sensitivity. PC can therefore be interpreted as a measure of spatial efficiency and risk sensitivity.

Robustness to Calibration-Test Split Ratios. To assess the robustness of CLS under varying calibration data availability, we evaluate its performance using different calibration-test split ratios. As shown in Figure 5b, CLS maintains strict control of the test-time ECR across all proportions, even when only 10% of the data is reserved for calibration. These results highlight a critical advantage of CLS: statistical guarantees remain valid even in imbalanced calibration settings—a desirable trait for scalable and trustworthy deployment of risk-aware medical AI systems.

4 Conclusion

In this paper, we propose CLS, a risk-controlled lesion segmentation framework that constructs an FNR-aware nonconformity score and derives statistically rigorous decision thresholds via conformal calibration. Unlike heuristic thresholding, CLS consistently enforces test-time FNR control while maintaining low FPR across diverse 3D-LS benchmarks, substantially reducing under-segmentation and complete lesion omission in safety-critical clinical settings. Beyond reliabil-

ity, we introduce *prediction compactness*, an interpretable auxiliary metric for quantifying spatial precision and uncertainty under formal risk constraints. We further show that CLS remains effective with limited calibration data. Overall, CLS provides a principled and practical foundation for uncertainty-aware segmentation with statistical guarantees, bridging theoretical rigor and real-world clinical reliability. We note that the exchangeability assumption underlying CLS may be weakened by scanner or protocol variations, temporal drift, and cross-institutional deployment; future work should incorporate domain-shift detection or site-specific recalibration to maintain validity under such distribution changes.

Acknowledgments. The paper was supported by Noncommunicable Chronic Diseases-National Science and Technology Major Project (2025ZD0551300, 2025ZD0551302).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Angelopoulos, A., Bates, S., Malik, J., Jordan, M.I.: Uncertainty sets for image classifiers using conformal prediction. arXiv preprint arXiv:2009.14193 (2020)
2. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021)
3. Antonelli, M., Reinke, A., Bakas, S., et al.: The medical segmentation decathlon. Nature Communications **13**(1), 4128 (2022)
4. Armato III, S.G., McLennan, G., Bidaut, L., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical Physics **38**(2), 915–931 (2011)
5. Bates, S., Angelopoulos, A., Lei, L., et al.: Distribution-free, risk-controlling prediction sets. Journal of the ACM **68**(6), 1–34 (2021)
6. Bilic, P., Christ, P., Li, H.B., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
7. Chen, J., Liu, Y., Wei, S., et al.: A survey on deep learning in medical image registration: New technologies, uncertainty, evaluation metrics, and beyond. Medical Image Analysis **100**, 103385 (2025)
8. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
9. de Grauw, M., Scholten, E.T., Smit, E.J., et al.: The uls23 challenge: A baseline model and benchmark dataset for 3d universal lesion segmentation in computed tomography. Medical Image Analysis **102**, 103525 (2025)
10. Hatamizadeh, A., Nath, V., Tang, Y., et al.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. pp. 272–284 (2021)
11. Hatamizadeh, A., Tang, Y., Nath, V., et al.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
12. He, Y., Guo, P., Tang, Y., et al.: Vista3d: A unified segmentation foundation model for 3d medical imaging. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20863–20873 (2025)

13. Heller, N., Isensee, F., Trofimova, D., et al.: The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. arXiv preprint arXiv:2307.01984 (2023)
14. Isensee, F., Jaeger, P.F., Kohl, S.A., et al.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
15. Jalalifar, S.A., Soliman, H., Sahgal, A., Sadeghi-Naini, A.: Impact of tumour segmentation accuracy on efficacy of quantitative mri biomarkers of radiotherapy outcome in brain metastasis. Cancers **14**(20), 5133 (2022)
16. Korhonen, K.E., Zuckerman, S.P., Weinstein, S.P., et al.: Breast mri: false-negative results and missed opportunities. Radiographics **41**(3), 645–664 (2021)
17. Luo, X., Yang, Y., Yin, S., Li, H., et al.: False-negative and false-positive outcomes of computer-aided detection on brain metastasis: Secondary analysis of a multicenter, multireader study. Neuro-Oncology **25**(3), 544–556 (2023)
18. Moglia, A., Cavicchioli, M., Mainardi, L., Cerveri, P.: Deep learning for pancreas segmentation on computed tomography: a systematic review. Artificial Intelligence Review **58**(8), 220 (2025)
19. Ni, G., Cao, K., Qin, X., Zeng, X., et al.: Advanced 3d retinal lesion segmentation using channel-spatial attention-guided multi-scale feature aggregation. Biomedical Optics Express **16**(5), 2093–2110 (2025)
20. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A.: Inductive confidence machines for regression. In: European Conference on Machine Learning. pp. 345–356 (2002)
21. Roth, H.R., Lu, L., Seff, A., Cherry, K.M., et al.: A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 520–527 (2014)
22. Sun, Y., Wang, L., Li, G., Lin, W., Wang, L.: A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. Nature Biomedical Engineering **9**(4), 521–538 (2025)
23. Wang, H., Guo, S., Ye, J., Deng, Z., et al.: Sam-med3d: towards general-purpose segmentation models for volumetric medical images. arXiv preprint arXiv:2310.15161 (2023)
24. Wang, Q., Geng, T., Wang, Z., Wang, T., Fu, B., Zheng, F.: Sample then identify: A general framework for risk control and assessment in multimodal large language models. arXiv preprint arXiv:2410.08174 (2024)
25. Wang, Z., Duan, J., Cheng, L., Zhang, Y., Wang, Q., Shi, X., Xu, K., Shen, H.T., Zhu, X.: Conu: Conformal uncertainty in large language models with correctness coverage guarantees. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 6886–6898 (2024)
26. Wang, Z., Wang, Q., Zhang, Y., Chen, T., et al.: Sconu: Selective conformal uncertainty in large language models. arXiv preprint arXiv:2504.14154 (2025)
27. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., et al.: Medical sam adapter: Adapting segment anything model for medical image segmentation. Medical Image Analysis **102**, 103547 (2025)
28. Ye, F., Yang, M., Pang, J., Wang, L., et al.: Benchmarking llms via uncertainty quantification. Advances in Neural Information Processing Systems **37**, 15356–15385 (2024)
29. Zhao, L., Wang, T., Chen, Y., Zhang, X., et al.: A novel framework for segmentation of small targets in medical images. Scientific Reports **15**(1), 9924 (2025)