CancerVerse: A Fully Open Longitudinal and Multimodal Dataset for Multicancer Screening

Wenxuan Li[1†], Pedro R. A. S. Bassi[1†], Xinze Zhou[1], Qi Chen[1], Jakob Wasserthal[2], Ibrahim Hamamci[3], Sezgin ER[3], Bjoern Menze[3], Gulhan Ertan Akan[4], Szymon Płotka[5], Jakub Prządo[6], Yucheng Tang[7], Daguang Xu[7], Arkadiusz Sitek[8], Kang Wang[9], Yang Yang[9], Alan L. Yuille[1], and Zongwei Zhou[1,10⋆]

[1] Johns Hopkins University
[2] University Hospital Basel
[3] University of Zurich
[4] Istanbul Medipol University
[5] Jagiellonian University
[6] Warmian-Masurian Cancer Center
[7] Nivida
[8] Harvard University
[9] University of California, San Francisco
[10] Johns Hopkins Medicine

Abstract. Cancer screening can greatly benefit from *longitudinal* and *multimodal* data. However, most public datasets are single time-point and image-only, lacking longitudinal follow-up, reports, pathology, and clinical variables. We present **CancerVerse**, the first large-scale, open-source longitudinal and multimodal dataset for multicancer screening. It contains **23,742** three-dimensional CT scans from **9,316** patients and provides voxel-wise tumor annotations for **13 cancer types** in the pelvis, abdomen, and chest regions. Each cancer patient has 1–19 longitudinal CT scans (median: 2), with every scan paired with radiology reports, pathology results, clinical variables, and laboratory tests, enabling joint modeling of imaging and non-imaging information. The dataset includes a substantial cohort of 2,498 healthy patients with at least 1 year of clinical follow-up, allowing realistic estimation of screening *specificity*. In this paper, we formulate screening as tumor detection/segmentation and evaluate performance at fixed false positives (FP) per scan to reflect real screening constraints. Experiments demonstrate that models trained on longitudinal and multimodal data systematically outperform those trained on single time-point, image-only data, achieving a **9%** absolute gain in sensitivity at 0.5 FP/scan. These results highlight the importance of modeling the full longitudinal patient trajectory of multimodal data, rather than isolated imaging snapshots, for effective cancer screening.

Keywords: cancer screening · longitudinal dataset · multimodal dataset · voxel-wise annotation · radiology reports · pathology · clinical data.

⋆ Correspondence to: Zongwei Zhou (zzhou82@jh.edu)

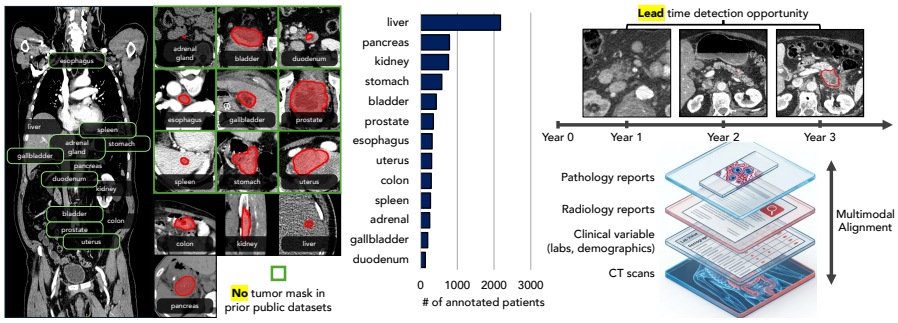

Fig. 1. CancerVerse is an open, longitudinal, and multimodal dataset for multicancer screening with three key features: *(i)* multiple CT scans per patient over time, *(ii)* paired radiology reports and clinical variables, and *(iii)* a verified healthy cohort with follow-up for realistic false-positive measurement. The dataset includes **23,742** CT scans from **9,316** patients with voxel-wise tumor annotations for **13 cancer types**. Each cancer patient has **1–19** longitudinal scans (median: 2) paired with radiology reports, pathology results, and laboratory tests.

1 Introduction

Early cancer detection from computed tomography (CT) is constrained by the lack of open datasets that reflect how cancers present, progress, and are documented in routine clinical care. Current public CT datasets are largely single time-point, image-only collections assembled around confirmed tumors [5,10,15] (therefore often in late stages). Most focus on a single organ and include only a limited number of cancer types, which restricts their use for developing and benchmarking multicancer imaging systems. In addition, they lack longitudinal trajectories, follow-up confirming healthy status, paired image and non-imaging information, and voxel-wise multicancer annotations. As a result, models trained on these datasets are primarily optimized for segmentation or diagnostic classification at the time of presentation; while they can be applied to screening scenarios, their performance—particularly for early cancer detection—may be limited [2,18], and they are not typically evaluated under screening conditions where temporal evolution, multimodal evidence, and fixed false positives per scan determine practical utility [4,25].

To address this gap, we introduce CancerVerse, a fully open, longitudinal, and multimodal dataset designed explicitly for multicancer screening. We formulate screening as tumor detection and segmentation, and evaluate performance at fixed false positives (FP) per scan to reflect practical screening constraints rather than retrospective diagnostic accuracy [14,16,23], making two contributions:

1. **A fully open longitudinal and multimodal dataset.** We create and release CancerVerse, a large-scale dataset designed for multicancer screening that integrates longitudinal CT scans with radiology/pathology reports, clinical variables, and laboratory tests. It includes comprehensive voxel-wise tu-

mor annotations across **13** cancer types (exampled in Fig. 1), including nine cancer types for which no prior public dataset provides related CT scans or voxel-wise tumor annotations, and a verified cohort of healthy individuals with >1 year follow-up for realistic specificity estimation.

2. **A comprehensive benchmark demonstrating the value of longitudinal and multimodal data.** Our experiments provide strong evidence that models trained on longitudinal and multimodal data systematically outperform those trained on single time-point, image-only data, achieving a **9%** absolute gain in sensitivity at both 0.25 and 0.5 FP/scan (Fig. 2). Specifically, our model achieved a sensitivity/specificity/DSC of **76**/**86**/**51%** for detecting and segmenting 13 cancer types (Table 1; Fig. 3). In external validation on 3,035 CT scans from three regional datasets (N. California, E. Coast, and E. Asia), detection/segmentation improves sensitivity/specificity by **2.6**/**1.4%** (Fig. 4), confirming generalization across diverse populations.

The CancerVerse dataset will be openly released for academic use under a CC BY-NC 4.0 license, with commercial use available by permission. We provide documentation, a standardized data format, and baseline tasks to enable rapid model development and comparison. Beyond the dataset, we will also release all trained segmentation backbones from our benchmarks. These will be the ***first*** set of public models that can detect and segment cancer in 13 organs.

Related work. Compared to existing public datasets that are typically single time-point, image-only, and organ- or tumor-specific (e.g., KiTS [10], LiTS [5], PanTS [15], BraTS [19] MSD [1]), CancerVerse is designed for screening by providing longitudinal trajectories, a verified healthy follow-up cohort for measuring specificity/false positives, paired imaging and non-imaging information, and voxel-wise multicancer annotations. Although some existing datasets include partial combinations of these components [3,8,9,13], none integrate all of them.

2 The CancerVerse Dataset

2.1 Cohort Construction and Longitudinal Structure

CancerVerse contains 9,316 patients and 23,742 pelvic, abdominal, and thoracic CT scans. The data were collected from multiple institutions in Western Europe and West Asia under appropriate IRB/ethics approvals. The acquisition time window spans 2012–2025. Each patient is represented by a time-ordered sequence of CT scans, with a median of 2 scans per patient and a median follow-up duration of 1.17 years. We include two complementary cohorts.

First, the **cancer-positive cohort** consists of patients with confirmed malignancies across 13 cancer types. For these patients, we define the *index date* as the date of pathological confirmation or, when pathology is unavailable, the date of first definitive diagnostic imaging. Critically, we retain not only the index scan but also all prior scans from the same patient when available. This temporal structure enables evaluation of screening performance at various lead times

before diagnosis—directly measuring whether models can detect cancer earlier than clinical presentation.

Second, the **verified healthy cohort** consists of patients with no cancer diagnosis and documented follow-up confirming healthy status for more than 1 year after the last included CT. This cohort is essential for screening evaluation because specificity measured on *"non-cancer scans near diagnosis"* substantially overestimates real-world performance. The verified healthy cohort provides a realistic substrate for measuring false positives per scan.

To prevent leakage across time, we split the dataset strictly at the patient level. We define training and test sets with no patient overlap: the ***official* training set** ($N_{\text{patients}} = 7,452$; $N_{\text{scans}} = 19,288$) and the ***official* test set** ($N_{\text{patients}} = 1,864$; $N_{\text{scans}} = 4,454$). Demographics (age and sex) are balanced between two splits to ensure similar distributions and reduce demographic bias.

2.2 Image Standarization and Tumor Annotation

All CT scans are provided in a standardized volumetric format with harmonized orientation and voxel spacing metadata. We release the original acquisition parameters (e.g., slice thickness, pixel spacing, reconstruction kernel when available) to enable robustness studies under protocol variation. Minimal preprocessing is applied to preserve realism; we provide standard intensity windowing (e.g., HU clipping to [-1000, 1000]) utilities to facilitate reproducible model training while allowing researchers to explore alternative pipelines.

A defining feature of CancerVerse is comprehensive voxel-wise tumor annotation across 13 cancer types. Annotation was performed using a multi-stage process designed to ensure consistency and reduce noise. Briefly, initial masks were produced by 28 radiologist residents using MONAI Label [6], followed by independent review by 8 board-certified radiologists. Discrepancies were resolved through adjudication, with particular attention to ambiguous boundaries and multifocal disease.

2.3 Multimodal Alignment: Reports and Clinical Variables

To support multimodal screening models, each CT scan is paired with its corresponding radiology report and structured clinical variables available at the time of imaging (e.g., demographics and selected EHR-derived fields). This alignment enables multiple research directions: multimodal fusion for detection, report-grounded learning, and evaluation of how non-imaging context affects false positive rates. Importantly, all non-imaging variables are time-stamped and linked to the imaging encounter to avoid inadvertent use of post-diagnostic information.

3 Longitudinal & Multimodal Model and Benchmark

The longitudinal & multimodal model. We combine two established techniques: temporal aggregation from prior work on longitudinal tumor detection

Table 1. Internal evaluation of multicancer detection. Sensitivity and specificity across 13 cancer types on the official test set of CancerVerse. We compare ULS models (public multicancer models), segmentation backbones (CancerVerse-trained), and our longitudinal & multimodal model. ULS models are evaluated only on cancers with existing public masks (liver, kidney, pancreas, colon). The longitudinal & multimodal model achieves the best average sensitivity/specificity of 76.1%/89.7% across all types.

tumors **w/** public masks	liver		kidney		pancreas		colon		**average**	
model & benchmark	sen.	spec.	sen.	spec.	sen.	spec.	sen.	spec.	sen.	spec.
Universal lesion segmentation (ULS) models validated on CancerVerse										
ULS Model 400 [8]	52.2	72.5	42.8	76.9	44.7	69.8	46.5	76.1	46.6	73.9
ULS Model 901 [8]	49.4	73.6	44.5	76.4	48.1	70.9	44.6	75.7	46.7	74.2
ULS+† [24]	53.8	78.9	48.6	72.7	45.2	70.3	50.1	75.5	49.4	74.3
Widely-adopted segmentation backbones developed and validated on CancerVerse										
SegResNet [20]	71.7	87.3	70.5	86.9	73.2	87.4	67.1	85.8	70.6	86.9
SwinUNETR [22]	72.1	83.4	70.8	84.9	73.5	85.6	65.2	82.3	70.4	84.1
Universal Model [17,18]	70.5	86.8	71.2	85.4	72.8	87.3	65.4	84.7	70.0	86.1
MedFormer [7]	72.6	87.3	73.8	88.7	74.1	89.2	67.9	83.4	72.1	87.2
nnU-Net [11]	73.2	88.5	72.4	87.1	75.9	88.9	67.5	86.0	72.3	87.6
longitudinal multimodal	77.5	90.2	76.1	90.6	78.9	91.5	71.8	86.3	76.1	89.7
tumors **w/o** public masks	adrenal		bladder		duodenum		esophagus		gallbladder	
SegResNet [20]	79.1	77.6	61.2	83.5	70.4	74.1	68.9	82.8	76.8	78.4
SwinUNETR [22]	71.4	81.2	63.7	74.6	62.9	83.4	69.8	75.1	70.6	78.3
Universal Model [17,18]	70.6	82.7	62.9	77.6	60.7	85.4	68.2	76.8	69.4	80.1
MedFormer [7]	76.5	79.4	62.4	80.1	66.7	77.2	71.3	79.8	74.1	80.3
nnU-Net [11]	72.9	83.8	65.7	78.6	63.8	84.9	70.4	77.5	71.9	81.2
longitudinal multimodal	75.8	84.4	68.9	82.3	66.7	85.6	73.6	80.9	74.2	83.1
tumors **w/o** public masks	spleen		stomach		prostate		uterus		**average**	
SegResNet [20]	82.7	86.2	62.5	76.9	75.9	81.7	76.6	81.3	72.6	80.3
SwinUNETR [22]	80.2	84.6	68.1	82.0	77.3	84.5	78.4	80.1	71.4	80.4
Universal Model [17,18]	78.5	86.9	67.1	81.2	76.0	84.5	77.8	75.6	70.1	81.2
MedFormer [7]	79.6	86.9	64.8	75.9	77.1	83.7	79.4	83.1	72.4	80.7
nnU-Net [11]	81.4	86.3	69.8	82.4	78.6	85.1	80.2	81.0	72.7	82.3
longitudinal multimodal	84.6	87.5	73.7	85.8	82.4	88.7	83.3	84.2	75.9	84.7

†Top-1 on the official Leaderboard of Universal Lesion Segmentation (ULS) Challenge.

[21] with report-based supervision from multimodal learning [2]. Specifically, we encode the longitudinal CT scans and fuse features across time to highlight changes between consecutive scans, instead of analyzing each scan independently. During training, we also use radiology reports to supervise the predicted tumor masks by extracting tumor attributes such as location and size and adding loss terms that reduce both false positives and false negatives beyond standard mask supervision. Together, longitudinal fusion helps detect subtle early lesions by modeling temporal change, and report supervision provides clinically grounded guidance that improves specificity when voxel-wise annotations are limited.

3.1 Comparing with state-of-the-art baselines and radiologists

Tumor detection and segmentation. We evaluated four approaches: (i) *Image-only*: nnU-Net on single CT scans [12]; (ii) *Longitudinal*: nnU-Net with temporal aggregation across patient scans [21]; (iii) *Multimodal*: nnU-Net incorpo-

rating radiology reports, pathology, and clinical variables [2]; and (iv) *Longitudinal & Multimodal*: our full model combining temporal context and non-imaging data. We additionally benchmark five segmentation backbones [7,11,17,20,22] and public universal lesion segmentation (ULS) methods [8,24] as reference. The ULS models are evaluated only on cancer types with existing public voxel-wise annotations, consistent with their original training setting. The segmentation backbones are trained and evaluated on all 13 cancer types.

Radiologist performance. We assessed and reported radiologist performance on a subset of the test set. Three board-certified radiologists independently produced voxel-wise tumor annotations for the cancer scans, and we quantify inter-annotator agreement using the Dice similarity coefficient (DSC), averaged across reader pairs. This serves as a human reference and helps contextualize model performance for multicancer segmentation.

3.2 Benchmark Protocol

False positives per scan, not ROC curves. In screening, we measure performance differently than diagnostic imaging. Rather than optimizing for a single threshold, we fix the number of false alarms (typically 0.25–1.0 per scan) and measure how many cancers we catch. This matters because in screening, false positives are costly—they lead to unnecessary follow-up imaging and patient anxiety. We evaluate sensitivity (how many cancers we detect) at these fixed false positive rates, a standard approach called free-response operating characteristic (FROC) in cancer screening.

Our evaluation metrics. We use sensitivity, specificity, and sensitivity at fixed false positives per scan (0.25 and 0.5 FP/scan) for the detection task, as well as Dice similarity coefficient (DSC) for the segmentation task.

External validation datasets. In addition to internal validation on the official test split of CancerVerse, we also provide external validation from three regional data resources. (1) *N. California* ($N_{\text{patients}} = 705$; $N_{\text{scans}} = 1,271$). (2) *E. Coast* ($N_{\text{patients}} = 366$; $N_{\text{scans}} = 870$). (3) *E. Asia* ($N_{\text{patients}} = 612$; $N_{\text{scans}} = 894$). All these three datasets, consisting of 3,035 CT scans, offer broad demographic coverage (age, sex, race) and diverse protocol settings (arterial/venous/non-contrast phases; heterogeneous acquisition parameters.

4 Results

4.1 Longitudinal & multimodal improves tumor *detection*

Table 1 shows that the longitudinal & multimodal model achieves 76.0% sensitivity and 86.2% specificity on the test set, demonstrating both reliable cancer detection and effective control of false positives on healthy follow-up scans. Fig. 2 shows the screening FROC curves for 13 cancer types in CancerVerse. Across operating points, longitudinal & multimodal achieves the best detection performance, with an absolute sensitivity gain of 8.6% at 0.25 FP/scan and 9.0%

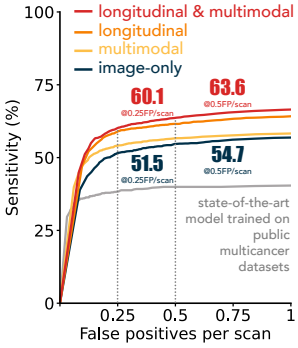

Fig. 2. Free-response receiver operating characteristic (FROC) curve for multicancer detection. Four training strategies are compared: *(i)* image-only (single CT) [11], *(ii)* longitudinal (multiple CT scans per patient) [21], *(iii)* multimodal (single CT with paired non-imaging data) [2], and *(iv)* longitudinal & multimodal (full patient trajectory). Longitudinal & multimodal models (red) achieve +**9%** sensitivity improvement over image-only baselines at both 0.25 and 0.5 FP/scan. Gray curve: the state-of-the-art model developed on existing public datasets [24].

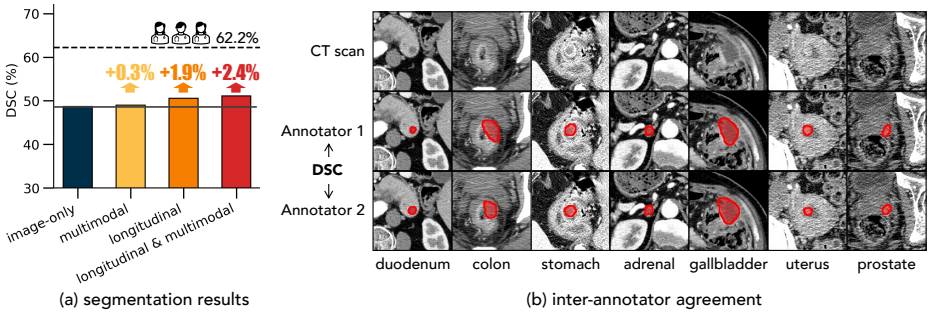

Fig. 3. Internal evaluation of multicancer segmentation. *(a)* Longitudinal & multimodal model achieves the highest DSC (**51.1%**), improving 2.4% over image-only baseline. Inter-annotator agreement among radiologists is **62.2%** DSC, reflecting the inherent difficulty of precise tumor annotation on CT. *(b)* Representative radiologist annotations for seven cancer types with subtle, ambiguous boundaries. Variability between annotators provides realistic reference for achievable performance. Incorporating longitudinal scans and clinical data improves spatial consistency of tumor localization.

at 0.5 FP/scan compared with the image-only baseline. Longitudinal-only and multimodal-only training each improves over image-only, indicating that temporal context and non-imaging evidence provide complementary benefits, while their combination yields the largest gains under low-FP constraints.

4.2 Longitudinal & multimodal improves tumor *segmentation*

Fig. 3 shows that the longitudinal & multimodal model achieves the highest mean Dice similarity coefficient (DSC) of 51.1%, improving spatial consistency of tumor localization. Importantly, inter-annotator agreement among three board-certified radiologists on the same tumors also achieves only 62.2% DSC (Fig. 3), reflecting the inherent difficulty of defining precise tumor boundaries on CT. Cancer margins are often subtle and anatomically ambiguous, introducing variability in expert radiologist segmentations. The model achieves 51.1% DSC, representing competitive performance relative to radiologists. longitudinal & multi-

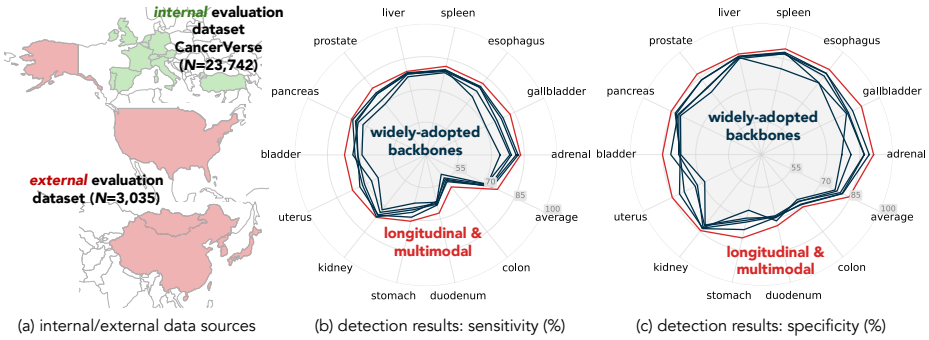

Fig. 4. External evaluation of multicancer detection. *(a)* Geographic distribution of CancerVerse (*W. Europe* and *W. Asia*) and three external cohorts (N. America and E. Asia). *(b–c)* Sensitivity and specificity across cancer types on external datasets. The longitudinal & multimodal model (red) consistently outperforms image-only baselines (dark gray) across regions and populations, demonstrating strong generalization.

modal model improves spatial consistency and reduces spurious detections compared to image-only baselines, particularly for subtle lesions.

4.3 Gains generalize across *external* regions and institutions

We evaluate generalization on 3,035 CT scans from three external regional datasets spanning *N. California*, *E. Coast*, and *E. Asia*. The models were trained on CancerVerse, which is collected from hospitals in *W. Europe* and *W. Asia*. As shown in Fig. 4, longitudinal & multimodal model improves sensitivity and specificity by 2.6% and 1.4%, respectively, relative to image-only training. These results suggest that combining temporal evidence with non-imaging context improves robustness beyond the CancerVerse training distribution, even when acquisition protocols and patient populations differ.

5 Discussion and Conclusion

The key insight from this work is that *screening models benefit substantially from longitudinal trajectories and multimodal data*. Existing multicancer datasets focus on single time-point diagnosis, but screening operates fundamentally differently: at low prevalence with fixed false-positive constraints. By aligning CT scans temporally and pairing them with radiology reports, pathology results, and clinical variables, we demonstrate a 9% absolute sensitivity gain at 0.5 FP/scan compared to image-only baselines. This gain generalizes across three independent external cohorts (*N. California*, *E. Coast*, *E. Asia*), suggesting that longitudinal and multimodal modeling is a practical and effective strategy for improving early cancer detection.

CancerVerse also reveals that voxel-wise tumor annotation on CT inherently involves ambiguity; inter-radiologist agreement reaches only 62% DSC, contextualizing our model's 51% DSC performance. This observation motivates uncertainty-aware model outputs and human-in-the-loop workflows rather than pursuit of unrealistic pixel-level perfection.

Key limitations include: *(i)* CT-centric focus, *(ii)* incomplete multimodal data availability, and *(iii)* the need for prospective validation. Despite these constraints, our findings establish a clear principle: screening research should move beyond isolated imaging snapshots toward modeling full patient trajectories that integrate both imaging and non-imaging information.

Acknowledgments. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the National Institutes of Health (NIH) under Award Number R01EB037669. We would like to thank the Johns Hopkins Research IT team in IT@JH for their support and infrastructure resources where some of these analyses were conducted; especially DISCOVERY HPC. We thank Jaimie Patterson for writing a news article about this project. Paper content is covered by patents pending.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., et al.: The medical segmentation decathlon. arXiv preprint arXiv:2106.05735 (2021)
2. Bassi, P.R., Li, W., Chen, J., Zhu, Z., Lin, T., Decherchi, S., Cavalli, A., Wang, K., Yang, Y., Yuille, A.L., Zhou, Z.: Learning segmentation from radiology reports. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 305–315. Springer (2025), https://github.com/MrGiovanni/R-Super
3. Bassi, P.R., Yavuz, M.C., Hamamci, I.E., Er, S., Chen, X., Li, W., Menze, B., Decherchi, S., Cavalli, A., Wang, K., Yang, Y., Yuille, A., Zhou, Z.: Radgpt: Constructing 3d image-text tumor datasets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23720–23730 (2025), https://github.com/MrGiovanni/RadGPT
4. Bassi, P.R., Zhou, X., Li, W., Płotka, S., Chen, J., Chen, Q., Zhu, Z., Prządo, J., Hamamci, I.E., Er, S., Chen, X., Yavuz, M.C., Chou, Y.C., Lin, T., Wang, K., Tang, Y., Cwikla, J.B., Decherchi, S., Cavalli, A., Yang, Y., Yuille, A.L., Zhou, Z.: Scaling artificial intelligence for multi-tumor early detection with more reports, fewer masks. arXiv preprint arXiv:2510.14803 (2025), https://github.com/MrGiovanni/R-Super
5. Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.W., Han, X., Heng, P.A., Hesser, J., et al.: The liver tumor segmentation benchmark (lits). arXiv preprint arXiv:1901.04056 (2019)
6. Diaz-Pinto, A., Alle, S., Nath, V., Tang, Y., Ihsani, A., Asad, M., Pérez-García, F., Mehta, P., Li, W., Flores, M., et al.: Monai label: A framework for ai-assisted interactive labeling of 3d medical images. Medical Image Analysis **95**, 103207 (2024)
7. Gao, Y., Zhou, M., Liu, D., Yan, Z., Zhang, S., Metaxas, D.N.: A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. arXiv preprint arXiv:2203.00131 (2022)

8. de Grauw, M., Scholten, E.T., Smit, E.J., Rutten, M.J., Prokop, M., van Ginneken, B., Hering, A.: The uls23 challenge: A baseline model and benchmark dataset for 3d universal lesion segmentation in computed tomography. Medical image analysis **102**, 103525 (2025)
9. Hamamci, I.E., Er, S., Menze, B.: Ct2rep: Automated radiology report generation for 3d medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 476–486. Springer (2024)
10. Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
12. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. arXiv preprint arXiv:2404.09556 (2024)
13. Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., et al.: Mimic-iv, a freely accessible electronic health record dataset. Scientific data **10**(1), 1 (2023)
14. Kashyap, M., Wang, X., Panjwani, N., Hasan, M., Zhang, Q., Huang, C., Bush, K., Chin, A., Vitzthum, L.K., Dong, P., et al.: Automated deep learning–based detection and segmentation of lung tumors at ct imaging. Radiology **314**(1), e233029 (2025)
15. Li, W., Zhou, X., Chen, Q., Lin, T., Bassi, P.R., Chen, X., Ye, C., Zhu, Z., Ding, K., Li, H., et al.: Pants: The pancreatic tumor segmentation dataset. In: Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2025), https://github.com/MrGiovanni/PanTS
16. Link, K.E., Schnurman, Z., Liu, C., Kwon, Y.J., Jiang, L.Y., Nasir-Moin, M., Neifert, S., Alzate, J.D., Bernstein, K., Qu, T., et al.: Longitudinal deep neural networks for assessing metastatic brain cancer on a large open benchmark. Nature Communications **15**(1), 8170 (2024)
17. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., Landman, B.A., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21152–21164 (2023), https://github.com/ljwztc/CLIP-Driven-Universal-Model
18. Liu, J., Zhang, Y., Wang, K., Yavuz, M.C., Chen, X., Yuan, Y., Li, H., Yang, Y., Yuille, A., Tang, Y., Zhou, Z.: Universal and extensible language-vision models for organ segmentation and tumor detection from abdominal computed tomography. Medical Image Analysis p. 103226 (2024), https://github.com/ljwztc/CLIP-Driven-Universal-Model
19. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993 (2015)
20. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: International MICCAI brainlesion workshop. pp. 311–320. Springer (2018)
21. Rokuss, M.R., Kirchhoff, Y., Roy, S., Kovacs, B., Ulrich, C., Wald, T., Zenk, M., Denner, S., Isensee, F., Vollmuth, P., et al.: Longitudinal segmentation of ms lesions

via temporal difference weighting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 64–74. Springer (2024)
22. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740 (2022)
23. Wang, C., Shao, J., He, Y., Wu, J., Liu, X., Yang, L., Wei, Y., Zhou, X.S., Zhan, Y., Shi, F., et al.: Data-driven risk stratification and precision management of pulmonary nodules detected on chest computed tomography. Nature Medicine **30**(11), 3184–3195 (2024)
24. Weber, R., Rocholl, N., de Grauw, M., Prokop, M., Smit, E., Hering, A.: Uls+: Data-driven model adaptation enhances lesion segmentation. arXiv preprint arXiv:2601.02988 (2026), accepted at BVM 2026
25. Xia, Y., Yu, Q., Chu, L., Kawamoto, S., Park, S., Liu, F., Chen, J., Zhu, Z., Li, B., Zhou, Z., Yuille, A.L., Fishman, E.K., Hruban, R.H.: The felix project: Deep networks to detect pancreatic neoplasms. medRxiv (2022)