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

Wenxuan Li[1,2,3†], Pedro R. A. S. Bassi[1,2,3†], Xinze Zhou[1], Jakob Wasserthal[4♡], Ibrahim E. Hamamci[5♡], Sezgin Er[5♡], Bjoern Menze[5♡], Gulhan E. Akan[6♡], Szymon Płotka[7], Jakub Prządo[8], Qi Chen[1], Yucheng Tang[9], Daguang Xu[9], Arkadiusz Sitek[2,3], Kang Wang[10], Yang Yang[10], Alan L. Yuille[1], and Zongwei Zhou[1,11⋆]

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

[†] *Equal contribution* [♡] *Open data contributors*

Models & Data: https://huggingface.co/datasets/BodyMaps/CancerVerse

Abstract. Cancer screening benefits greatly from *longitudinal* and *multimodal* data. However, most public CT datasets are single time-point and image-only. They lack longitudinal follow-up, radiology reports, pathology results, and clinical variables. This makes them poorly suited to screening, where cancers must be detected early at a low rate of false positives per scan. We present **CancerVerse**, the first large-scale open-source longitudinal and multimodal dataset for multicancer screening. It contains **24,168** CT scans from **13,863** patients, with voxel-wise tumor annotations for **13 cancer types** in the pelvis, abdomen, and chest. Each cancer patient has 1 to 32 longitudinal CT scans, paired with radiology reports, pathology results, and clinical variables. A **8,799**-patient verified healthy cohort with more than one year of follow-up enables realistic specificity estimation. We formulate screening as tumor detection and segmentation, and evaluate sensitivity at fixed false positives (FP) per scan. Models trained on longitudinal and multimodal data outperform those trained on single time-point, image-only data, with a **9%** absolute sensitivity gain at 0.5 FP/scan. These results show that effective cancer screening depends on modeling the full longitudinal patient trajectory with multimodal data, rather than isolated imaging snapshots.

Keywords: cancer screening · longitudinal data · multimodal data

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

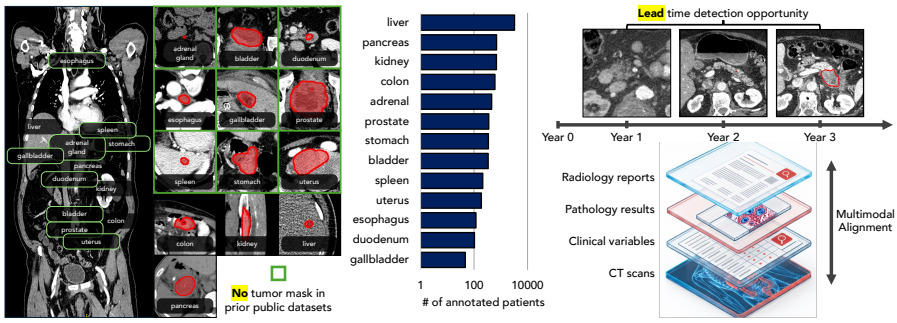

Fig. 1. CancerVerse is an open longitudinal and multimodal dataset for multicancer screening, with three key features: *(i)* multiple CT scans per patient over time; *(ii)* paired radiology reports, pathology results, and clinical variables; and *(iii)* a verified healthy cohort with follow-up for realistic false positive measurement. The dataset includes **24,168** CT scans from **13,863** patients with voxel-wise tumor annotations for **13 cancer types**. Each cancer patient has **1 to 32** longitudinal scans. Among those with follow-up, the median is three scans over 299 days.

1 Introduction

Early cancer screening from computed tomography (CT) is constrained by the lack of open datasets that present how cancers arise and progress in routine clinical care. Current public CT datasets fall short on four counts, and each gap directly limits what a screening model can learn. *Because* they are single time-point [6,14,23], models cannot learn how tumors change over time. *Because* they are image-only, models cannot combine imaging with non-imaging data such as radiology reports, pathology results, and clinical variables. *Because* they provide voxel-wise tumor annotations for only a few cancer types, models cannot deal with multiple types of cancer. *Because* they include no verified healthy cohort, a model's false positives on healthy scans cannot be reliably measured. Together, these gaps define what cancer screening requires: *detecting cancers early while keeping the number of false positives per scan low* [2,19,32].

To address these gaps, we introduce CancerVerse, a fully open longitudinal and multimodal dataset designed for multicancer screening. We measure how many cancers the model accurately detects, localizes, and segments at a fixed number of false positives (FP) per scan, since a screening tool is only useful when false positives stay low [18,24,31]. We make two contributions:

1. **A fully open longitudinal and multimodal dataset.** We create and release CancerVerse, a large-scale dataset for multicancer screening. It integrates longitudinal CT scans with radiology reports, pathology results, and clinical variables. It provides comprehensive voxel-wise tumor annotations across **13** cancer types (examples in Fig. 1). Seven of these cancer types have no prior public dataset with related CT scans or voxel-wise tumor annotations; AbdomenAtlas 2.0 [8] recently released annotations for esophagus

and uterus. CancerVerse also includes a verified cohort of healthy people with more than one year of follow-up for realistic specificity estimation.

2. **A comprehensive benchmark demonstrating the value of longitudinal and multimodal data.** Our experiments show that models trained on longitudinal and multimodal data systematically outperform those trained on single time-point, image-only data. They achieve absolute sensitivity gains of **8.6%** and **9.0%** at 0.25 and 0.5 FP/scan, respectively (Fig. 2). Our model reaches an average sensitivity, specificity, and DSC of **76.0%, 86.2%, and 51.1%**, respectively, across all 13 cancer types (Table 1; Fig. 3). On 3,035 external CT scans from three regional datasets (N. California, E. Coast, and E. Asia), our model improves sensitivity by **2.6%** and specificity by **1.4%** (Fig. 4). These results confirm generalization across diverse populations.

The CancerVerse dataset is openly released for academic use under a CC BY-NC-ND 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 have also released all trained segmentation backbones from our benchmarks. These mark the ***first*** set of public models that can detect, localize, and segment cancer in 13 organs.

Related work. Existing public datasets are typically single time-point, image-only, and organ- or tumor-specific (e.g., KiTS [14], LiTS [6], PanTS [23], BraTS [27], MSD [1]), or focused on multi-organ segmentation (e.g., AbdomenAtlas [21]). In contrast, CancerVerse is designed for screening. It provides longitudinal trajectories, a verified healthy follow-up cohort for measuring specificity and false positives, paired imaging and non-imaging data, and voxel-wise multicancer annotations. Some existing datasets include partial combinations of these components [5,22,12,13]. None integrate all of them.

Our broader vision follows MIMIC [16,17], the most comprehensive open medical dataset to date: its latest release links de-identified records for more than 360,000 patients across laboratory tests, medications, diagnoses, clinical notes, and chest radiographs with paired reports. The MIMIC dataset, however, is limited to 2D chest X-rays, with no 3D CT, no voxel-wise tumor annotations, and no multicancer coverage. CancerVerse brings this level of multimodal and longitudinal integration to 3D CT for multicancer screening. Among publicly released abdominal CT datasets, CancerVerse is one of only two *exceeding* 20,000 first-time-released scans, alongside Stanford Merlin [7], and the *only one* that combines voxel-wise tumor annotations across 13 cancer types with longitudinal trajectories, paired radiology reports and pathology results, clinical variables, and a verified healthy follow-up cohort.

2 The CancerVerse Dataset

Cohort construction and longitudinal structure. CancerVerse contains 13,863 patients and 24,168 pelvic, abdominal, and thoracic CT scans. The data

were collected from multiple institutions in Switzerland and Turkey under appropriate IRB (Req-2025-00557) and ethics approvals. The acquisition window spans 2010 to 2025. Each patient is represented by a time-ordered sequence of CT scans. Among those patients with follow-up, the median number of scans per patient is three, with a median follow-up duration of 299 days. 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. When pathology is unavailable, we use the date of first definitive diagnostic imaging. 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. It directly measures whether models can detect cancer earlier than clinical presentation.

Second, the **verified healthy cohort** contains 8,799 patients (63.5% of the dataset), none diagnosed with any of the 13 cancer types studied in this paper. The cohort includes documented follow-up confirming the absence of these cancers for more than one year after the last included CT scan. This cohort matters for evaluation. *Many prior studies measure specificity on a set of unusually clean scans. Such scans lack the everyday benign findings that cause false positives, so the reported specificity is often too optimistic.* Our healthy cohort pools three subgroups: patients with benign lesions only (47.6%), patients with malignancy outside the 13 cancer types of interest (32%), and cancer-free patients (20.4%).

To prevent leakage across time, we split the dataset strictly at the patient level. The ***official* training set** contains $N_{\text{patients}} = 11,091$. The ***official* test set** contains $N_{\text{patients}} = 2,772$. Demographics are balanced between the two splits to ensure similar distributions and reduce demographic bias [26].

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. Per-modality availability across the 13,863 patients is as follows: radiology reports 99.8%, pathology results 44.8% of all patients, demographics 80.4%, and clinical variables 74.7%. The full per-patient availability matrix has been released with the dataset.

Image standardization. All CT scans use a standard volumetric format with consistent orientation and voxel spacing. We release the original acquisition parameters when available, such as slice thickness, pixel spacing, and reconstruction kernel. These parameters enable robustness studies under protocol variation. We apply minimal preprocessing to preserve realism. We also provide standard intensity windowing, i.e., HU clipping to [-1000, 1000].

Multicancer annotation. A defining feature of CancerVerse is comprehensive voxel-wise tumor annotation across 13 cancer types. While many large-scale CT datasets accelerate annotation via semi-automated pipelines [21] or weak supervision for temporal and volumetric data [9], CancerVerse relies entirely on manual annotation: every cancerous lesion across the 13 organs was drawn by a human expert. No AI was used in the annotation pipeline, and no pseudo-

Table 1. Internal evaluation of multicancer detection. Sensitivity and specificity across 13 cancer types on the test set of CancerVerse. We compare widely-adopted segmentation backbones and our longitudinal and multimodal model. Averages are reported separately for the cancer types with and without public masks. Across all 13 cancer types, the longitudinal and multimodal model achieves the best sensitivity of 76.0% and specificity of 86.2%.

tumors **w/** public masks	liver		kidney		pancreas		colon		**average**	
model & benchmark	sen.	spec.	sen.	spec.	sen.	spec.	sen.	spec.	sen.	spec.
Widely-adopted segmentation backbones developed and validated on CancerVerse										
SegResNet [28]	71.7	87.3	70.5	86.9	73.2	87.4	67.1	85.8	70.6	86.9
SwinUNETR [30]	72.1	83.4	70.8	84.9	73.5	85.6	65.2	82.3	70.4	84.1
Universal Model [25]	70.5	86.8	71.2	85.4	72.8	87.3	65.4	84.7	70.0	86.1
MedFormer [11]	72.6	87.3	73.8	88.7	74.1	89.2	67.9	83.4	72.1	87.2
nnU-Net [15]	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 [28]	79.1	77.6	61.2	83.5	70.4	74.1	68.9	82.8	76.8	78.4
SwinUNETR [30]	71.4	81.2	63.7	74.6	62.9	83.4	69.8	75.1	70.6	78.3
Universal Model [25]	70.6	82.7	62.9	77.6	60.7	85.4	68.2	76.8	69.4	80.1
MedFormer [11]	76.5	79.4	62.4	80.1	66.7	77.2	71.3	79.8	74.1	80.3
nnU-Net [15]	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 [28]	82.7	86.2	62.5	76.9	75.9	81.7	76.6	81.3	72.6	80.3
SwinUNETR [30]	80.2	84.6	68.1	82.0	77.3	84.5	78.4	80.1	71.4	80.4
Universal Model [25]	78.5	86.9	67.1	81.2	76.0	84.5	77.8	75.6	70.1	81.2
MedFormer [11]	79.6	86.9	64.8	75.9	77.1	83.7	79.4	83.1	72.4	80.7
nnU-Net [15]	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

labels were used. The full annotation effort took 28 human experts more than 13 months, following the trusted-dataset development methodology of Scale-MAI [20]. Initial masks were produced by radiologist residents using MONAI Label [10] as the annotation interface (not for AI-generated pre-masks). Eight of these experts, all board-certified radiologists, then performed independent review. Discrepancies were resolved through adjudication, with particular attention to ambiguous boundaries and multifocal disease. As a quality assurance step, a subset of tumors was annotated independently by multiple experts to verify consistency, in line with recent label-quality assessment practices [4]; the resulting inter-expert Dice agreement is reported in Fig. 3a.

3 Longitudinal/Multimodal Modeling and Benchmarking

The longitudinal and multimodal model. We combine two established techniques. The first is temporal aggregation from prior work on longitudinal tumor detection [29]. The second is report-based supervision from multimodal learning [2]. Specifically, we encode the longitudinal CT scans and fuse features across time. This fusion highlights changes between consecutive scans, instead of analyzing each scan independently. To accommodate a variable number of scans per

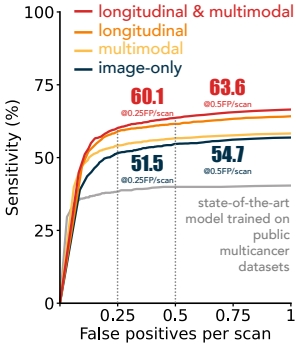

Fig. 2. Free-response receiver operating characteristic (FROC) curve for multicancer detection. Four training strategies are compared: *(i)* image-only (single CT) [15]; *(ii)* longitudinal (multiple CT scans per patient) [29]; *(iii)* multimodal (single CT with paired non-imaging data) [2]; and *(iv)* longitudinal and multimodal (full patient trajectory). Longitudinal and multimodal models (red) achieve absolute sensitivity gains of **+8.6%** and **+9.0%** over image-only baselines at 0.25 and 0.5 FP/scan, respectively. Gray curve: the state-of-the-art model developed on existing public datasets [12].

patient, a shared nnU-Net encoder processes each scan, followed by a temporal-difference layer across consecutive scans [29]; single-scan patients fall back to the image-only setting. During training, we also use radiology reports to supervise the predicted tumor masks. We extract tumor attributes such as location and size from the reports. Reports are encoded by Llama-3-70B with tumor-attribute auxiliary labels [2], clinical variables are encoded by a multilayer perceptron, and imaging and non-imaging features are fused via cross-attention. We then add 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. Report supervision provides clinically grounded guidance that improves specificity when voxel-wise annotations are limited.

Tumor detection and segmentation (Table 1; Fig. 2). We evaluate four approaches: (i) *Image-only*: nnU-Net on single CT scans [15]; (ii) *Longitudinal*: nnU-Net with temporal aggregation across patient scans [29]; (iii) *Multimodal*: nnU-Net incorporating radiology reports, pathology, and clinical variables [2]; and (iv) *Longitudinal and Multimodal (ours)*: the full model combining temporal context and non-imaging data. We also benchmark five segmentation backbones [11,15,25,28,30]. The segmentation backbones are trained and evaluated on all 13 cancer types. For a fair comparison, each baseline is reported at the operating point nearest the top-left of its ROC curve.

Radiologist performance (Fig. 3). We assess radiologist performance on a subset of the test set. Three board-certified radiologists independently produce voxel-wise tumor annotations for the cancer-positive scans. We quantify inter-annotator agreement using the Dice similarity coefficient (DSC), averaged across reader pairs. This serves as a human reference and contextualizes model performance for multicancer segmentation.

False positives per scan, not ROC curves. In screening, we measure performance differently than in diagnostic imaging. Rather than optimizing for a single threshold, we fix the number of false positives (typically 0.25 to 1.0 per scan). We then measure how many cancers we catch. False positives in screening 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

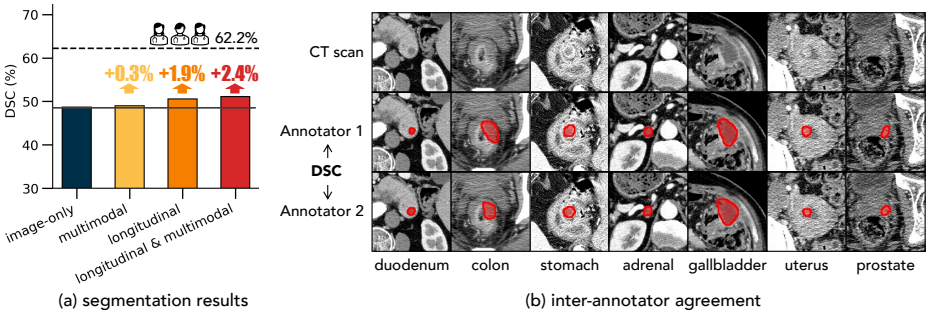

Fig. 3. Evaluation of multicancer segmentation. *(a)* Bars show segmentation DSC on the full test set for four training strategies; the longitudinal and multimodal model (red) achieves the highest mean DSC (**51.1%**), an improvement of 2.4% over the image-only baseline. The dashed line marks the inter-annotator agreement among radiologists (**62.2%** DSC), measured on a smaller subset of tumors independently annotated by three experts, since multi-expert annotation is time-consuming and expensive. On this same subset, the model reaches **50.8%** DSC, providing an apples-to-apples comparison with radiologists. The 62.2% inter-annotator DSC reflects 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 a realistic reference for achievable performance. Incorporating longitudinal scans and non-imaging data improves the spatial consistency of tumor localization.

rates. This is the standard free-response operating characteristic (FROC) formulation used in cancer screening: sensitivity is computed at the *lesion level*, where each true lesion is scored independently as detected or missed, and false positives are counted per scan. This is distinct from a patient-level aggregation in which any prediction in an image would be counted as a true positive. Choosing the right metric and operating point is increasingly recognized as central to medical AI evaluation [3,26].

Our evaluation metrics. For localization and segmentation, we use Dice similarity coefficient (DSC). For detection, we use sensitivity, specificity, and sensitivity at fixed false positives per scan (0.25 and 0.5 FP/scan). Negative samples for specificity and false positives per scan are drawn from the verified healthy cohort and from patients with a different cancer than the one being evaluated. This mirrors real screening conditions, in which a model evaluated on one cancer type encounters both healthy patients and patients with other diseases.

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

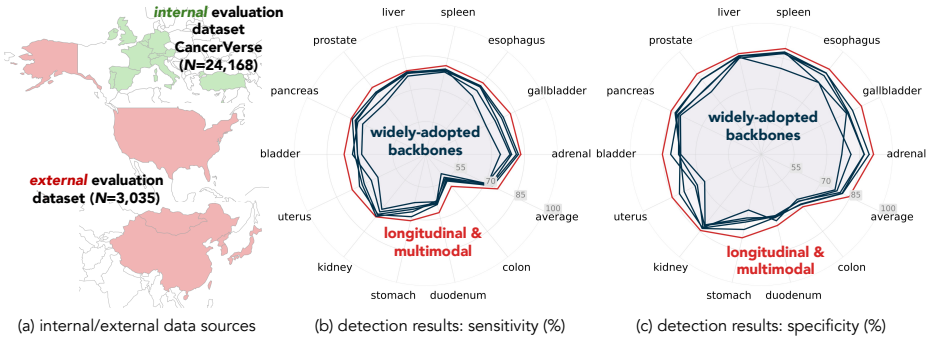

Fig. 4. External evaluation of multicancer detection. *(a)* Geographic distribution of CancerVerse (*Switzerland* and *Turkey*) and three external cohorts (*N. California*, *E. Coast*, and *E. Asia*). *(b, c)* Sensitivity and specificity across cancer types on external datasets. The longitudinal and multimodal model (red) consistently outperforms image-only baselines (dark gray; i.e., [28,30,25,11,15]) across regions and populations. This demonstrates strong generalization.

4 Results

Longitudinal/multimodal models improve tumor *detection*. Table 1 shows that the longitudinal and multimodal model achieves 76.0% sensitivity and 86.2% specificity, averaged across all 13 cancer types on the test set. This result demonstrates 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, the longitudinal and multimodal model achieves the best detection performance. It yields an absolute sensitivity gain of 8.6% at 0.25 FP/scan and 9.0% at 0.5 FP/scan compared with the image-only baseline. Longitudinal-only and multimodal-only training each improve over image-only training. This indicates that temporal context and non-imaging data provide complementary benefits. Their combination yields the largest gains under low-FP constraints.

Longitudinal/multimodal models improve tumor *segmentation*. Fig. 3 shows that the longitudinal and multimodal model achieves the highest mean DSC of 51.1% on the full test set. This means it locates tumors more consistently. Per-cancer DSC is 67% (liver), 64% (kidney), 48% (pancreas), 41% (colon), 52% (spleen), 45% (bladder), 46% (gallbladder), 55% (adrenal), 49% (esophagus), 38% (stomach), 39% (duodenum), 63% (uterus), and 57% (prostate). Scores are higher where organ boundaries are clearer. To compare against radiologists, three board-certified radiologists each marked a subset of tumors on their own. On this subset, the radiologists agree with each other at 62.2% DSC, and the model reaches 50.8%. The 62.2% inter-annotator agreement shows how hard it is to draw exact tumor edges on CT, where margins are often faint and unclear. The model's performance therefore comes close to radiologists on the same cases.

The longitudinal and multimodal model also predicts fewer false positives than image-only baselines, especially for faint lesions.

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

Conclusion. The key insight is that *screening models benefit substantially from longitudinal trajectories and multimodal data*. Existing multicancer datasets focus on single time-point diagnosis. Screening operates 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 over image-only baselines. This gain generalizes across three independent external cohorts (*N. California*, *E. Coast*, and *E. Asia*). These results suggest that longitudinal and multimodal modeling is a practical and effective strategy for screening.

Acknowledgments. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research, the National Institutes of Health (NIH) under Award Number R01EB037669 and National Science Center, Poland 2025/57/B/ST6/03100.

Disclosure of Interests. The authors declare no competing interests.

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: MICCAI. pp. 305–315. Springer (2025)
3. Bassi, P.R., Li, W., Tang, Y., Isensee, F., et al.: Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? NeurIPS **37**, 15184–15201 (2024)
4. Bassi, P.R., Wu, Q., Li, W., Decherchi, S., Cavalli, A., Yuille, A., Zhou, Z.: Label critic: Design data before models. In: ISBI. pp. 1–5. IEEE (2025)
5. Bassi, P.R., Yavuz, M.C., Hamamci, I.E., Er, S., Chen, X., Li, W., et al.: Radgpt: Constructing 3d image-text tumor datasets. In: ICCV. pp. 23720–23730 (2025)
6. 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)
7. Blankemeier, L., Kumar, A., Cohen, J.P., Liu, J., Liu, L., Van Veen, D., Gardezi, S.J.S., Yu, H., Paschali, M., Chen, Z., et al.: Merlin: a computed tomography vision–language foundation model and dataset. Nature pp. 1–11 (2026)

8. Chen, Q., Zhou, X., Liu, C., Chen, H., Li, W., et al.: Scaling tumor segmentation: Best lessons from real and synthetic data. In: ICCV. pp. 24001–24013 (2025)
9. Chou, Y.C., Li, B., Fan, D.P., Yuille, A., Zhou, Z.: Acquiring weak annotations for tumor localization in temporal and volumetric data. Machine Intelligence Research pp. 1–13 (2024)
10. 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)
11. 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)
12. 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)
13. 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)
14. Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., 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)
15. 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)
16. 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)
17. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., et al.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
18. Kashyap, M., Wang, X., Panjwani, N., Hasan, M., Zhang, Q., Huang, C., Bush, K., et al.: Automated deep learning–based detection and segmentation of lung tumors at ct imaging. Radiology **314**(1), e233029 (2025)
19. Li, W., Bassi, P.R.A.S., Wu, L., et al.: Early and prediagnostic detection of pancreatic cancer from computed tomography. arXiv preprint arXiv:2601.22134 (2026)
20. Li, W., Bassi, P.R., Lin, T., Chou, Y.C., Zhou, X., Tang, Y., Isensee, F., Wang, K., Chen, Q., Xu, X., Ye, J., Zhu, Z., et al.: Scalemai: Accelerating the development of trusted datasets and ai models. arXiv preprint arXiv:2501.03410 (2025)
21. Li, W., Qu, C., Chen, X., Bassi, P.R., et al.: Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. Medical Image Analysis p. 103285 (2024)
22. Li, W., Yuille, A., Zhou, Z.: How well do supervised 3d models transfer to medical imaging tasks? In: International Conference on Learning Representations (2024)
23. Li, W., Zhou, X., Chen, Q., Lin, T., Bassi, P.R., Chen, X., Ye, C., Zhu, Z., Ding, K., Li, H., Wang, K., Yang, Y., Tang, Y., Xu, D., Yuille, A.L., Zhou, Z.: Pants: The pancreatic tumor segmentation dataset. In: NeurIPS (2025)
24. 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)

25. 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: ICCV. pp. 21152–21164 (2023)
26. Lubonja, A., Bassi, P.R., Li, W., Qiao, H., Burns, R., Yuille, A.L., Zhou, Z.: Auditing significance, metric choice, and demographic fairness in medical ai challenges. arXiv preprint arXiv:2512.19091 (2025)
27. 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)
28. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: International MICCAI brainlesion workshop. pp. 311–320. Springer (2018)
29. 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: MICCAI. pp. 64–74. Springer (2024)
30. 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: CVPR. pp. 20730–20740 (2022)
31. 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)
32. Xia, Y., Yu, Q., Chu, L., Kawamoto, S., Park, S., Liu, F., Chen, J., Zhu, Z., Li, B., Zhou, Z., et al.: The felix project: Deep networks to detect pancreatic neoplasms. MedRxiv pp. 2022–09 (2022)