BENCHX: BENCHMARKING AI MODELS FOR CANCER DETECTION AND LOCALIZATION WITH DEMOGRAPHIC AND PROTOCOL BIASES

Anonymous authors
Paper under double-blind review

ABSTRACT

Artificial intelligence (AI) has achieved remarkable success in medical imaging, but it is widely recognized that these models often perform inconsistently across real-world clinical settings. Such inconsistencies occur when patient demographics and imaging protocols vary, for example, in detecting small tumors, analyzing scans from different contrast phases, or evaluating patients of different ages or sexes. To quantify these inconsistencies, we develop a large-scale, open benchmark of 101,822 CT scans that systematically evaluates 12 tumor-detection AI models across tumor size, location, patient subgroup, and imaging protocol. We leverage large language models (LLMs) to extract and organize subgroup information from clinical data, which makes the analysis both scalable and reproducible. Our benchmark reveals that current state-of-the-art AI models, optimized for average accuracy, perform poorly in rare or underrepresented subgroups, such as young, female African Americans. However, collecting sufficient annotated data for these rare cases is often impractical. The benchmark provide a foundation for building more reliable and robust AI models for tumor detection and highlighting the need for rigorous, subgroup-level evaluation in medical imaging and computer vision. Code, models, and data will be released upon acceptance. In total, 54,624 CT scans will be publicly accessible, supporting direct evaluation of demographic and protocol biases.

1 INTRODUCTION

Bias in data-driven algorithms has long been a recognized challenge. In the 1980s, studies in both medicine and computer vision showed that models trained on narrow populations often failed to generalize. Early face-recognition systems, such as *Eigenfaces* in the late 1980s and early 1990s (Sirovich & Kirby, 1987; Turk & Pentland, 1991; Phillips et al., 2000; Buolamwini & Gebru, 2018), relied mainly on images of adult men and performed poorly on women and other groups. Similar biases existed in biomedical research, where most clinical trials enrolled only men (Holdcroft, 2007; of Medicine, 1994), leading to diagnostic errors and prompting the NIH Revitalization Act of 1993 (Chen Jr et al., 2014). These early lessons underscore the need to test models across diverse populations—a principle especially crucial in medical imaging, where variations in patients and imaging protocols can strongly affect AI performance.

In the 2020s, while AI has greatly advanced tumor detection, offering more accurate and accessible cancer care worldwide (Xia et al., 2023; Bassi et al., 2025b; Chen et al., 2023), equitable performance across patient populations remains a big, unsolved challenge. Recent advances in large-scale medical imaging models demonstrate impressive accuracy when evaluated on single-institution datasets with homogenous patient cohorts (Asian population, for example, in (Cao et al., 2023) and (Hu et al., 2025)). However, there is a significant knowledge gap in how current models handle population and image acquisition protocol diversity.

The main bottlenecks are the lack of training data and evaluation benchmarks that explicitly account for demographic and acquisition variability. As a result, most models are implicitly biased toward well-represented groups (Bassi et al., 2024). This paper seeks to answer the critical question: *How can we build AI models that generalize robustly across patient demographics and imaging protocols*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

for tumor detection and localization? Existing solutions like domain adaptation partially address this issue, but they rely on narrow datasets and fail to assess group-wise performance biases in a systematic way (Jain et al., 2022; Yin et al., 2024; Hu et al., 2024).

To address these two bottlenecks, we construct a comprehensive dataset with detailed demographic and protocol factors. Specifically, we establish a benchmark, called **BenchX**, that measures model reliability across patient demographics (age, sex, race) and CT imaging protocols (non-contrast, arterial, venous, delayed). To build this benchmark at scale, we leverage large language models (LLMs) (Bai et al., 2023) to curate and organize data from over 450,000 CT scans from 112 global hospitals. LLMs can automatically read radiology reports and clinical notes quickly and accurately, summarizing key details such as patient characteristics and scan types. This automated process allows us to assemble a large and well-structured dataset of 101,822 CT scans with precisely matched patient demographics and imaging protocols. By balancing the representation of both common and less common subgroups, the resulting benchmark enables systematic evaluation of model robustness across clinically and demographically varied groups rather than relying on overall average accuracy. These annotations also enable single-factor analyses, allowing us to compare subgroup performance while controlling for measured confounders. Unlike traditional benchmarks such as DeepLesion (Yan et al., 2018), MSD (Antonelli et al., 2022), and FLARE (Ma et al., 2024b), which report only overall accuracy, BenchX reveals where models perform inconsistently across different patient demographics and imaging protocols, and examines large-scale tumor segmentation with more than 10× the number of test patients.

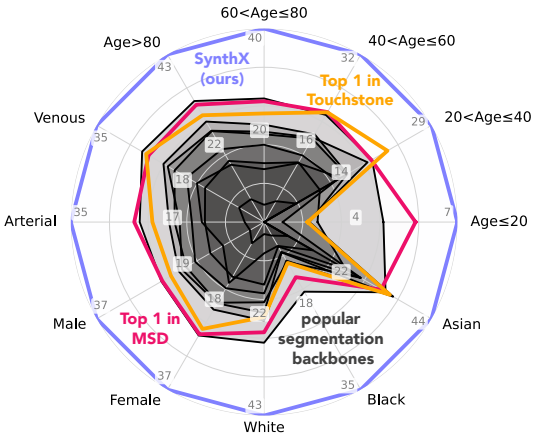
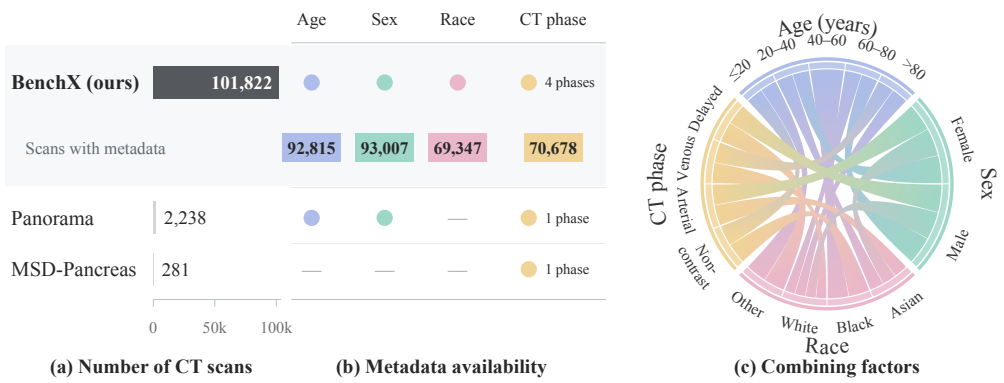


Figure 1: Existing tumor-detection models, including top leaderboard performers (Antonelli et al., 2021; Bassi et al., 2024), show inconsistent performance across patient demographics (age, race, sex) and imaging protocols (contrast phases). This benchmark (BenchX) systematically measures and exposes these performance gaps.

Our benchmark reveals clear and consistent patterns of performance degradation across patient demographics and imaging protocols. As shown in Figure 1, all widely used segmentation backbones, including the top models from MSD (Antonelli et al., 2022; Chen et al., 2025) and Touchstone (Bassi et al., 2024; Liu et al., 2024), lose substantial accuracy in young patients, minority racial groups, and non-venous CT phases. These gaps match the quantitative results across six external cohorts (see Figure 5). Arterial and non-contrast scans reduce F1 and sensitivity by 1530%, pediatric and young-adult patients show the lowest detection accuracy, and minority groups such as Black or Pacific Islander patients experience the largest drops (e.g., F1 as low as 2.5% on Merlin (Blankemeier et al., 2026)). Tumor characteristics further magnify these challenges: small tumors and pancreatic tail lesions consistently show the lowest performance across datasets. These results demonstrate that demographics, protocols, and tumor characteristics jointly drive the largest performance drops.

In summary, this paper makes **two key contributions**: **(1)** BenchX, a large-scale benchmark with fine-grained demographic and protocol annotations, enabling controlled comparisons to systematically expose fairness and domain-generalization gaps in tumor detection and localization models. **(2)** Five critical observations through comprehensive evaluation: (i) demographic disparities across age, sex, and race, (ii) performance degradation across CT imaging protocols, (iii) amplified challenges with small tumors and tumor location, (iv) consistent patterns across multiple state-of-the-art models, and (v) validation across six external cohorts demonstrating generalizability of these findings.

Code, trained models, and benchmark data will be released upon acceptance to promote transparent and equitable medical AI research. In total, 54,624 CT scans will be publicly accessible, supporting direct evaluation of demographic and protocol biases. The remaining 47,198 CT scans will be used for third-party model evaluation to support the research community.



(a) Number of CT scans **(b) Metadata availability** **(c) Combining factors**

Figure 2: **Dataset scale and subgroup coverage.** (a) CT scan counts in BenchX, MSD-Pancreas (Antonelli et al., 2021), and Panorama (Alves et al., 2026). (b) Metadata availability: dots denote available fields and dashes denote unreported fields. Highlighted values beneath the BenchX row count scans with each metadata field. (c) Age, sex, race, and CT phase can be combined to define finer subgroups among scans with the corresponding metadata. Ribbons connect all six factor pairs schematically. Full statistics are provided in Appendix B.

2 Related Work

Benchmarking AI robustness to biases. Bias has long been recognized in pedestrian detection, skin lesion analysis, and speech processing, where models trained on narrow populations tend to underperform on minority groups (Ktena et al., 2024; Lotter, 2024; Glocker et al., 2023). These findings highlight that data imbalance can lead to systematic errors unless models are evaluated and improved across diverse subgroups (Tian et al., 2023; Yang et al., 2024; Wang et al., 2024). Bias appears in tumor detection and localization as well, but existing studies remain limited and fragmented. Prior work reports that model accuracy drops for small tumors. In this paper, we take a step beyond average accuracy and perform the first systematic study of demographic and protocol biases in tumor detection. We build a large-scale, comprehensive benchmark that evaluates mainstream AI models across age, sex, race, tumor size, and CT phase.

Benchmarking AI for tumor detection and localization. Public benchmarks such as MSD (Antonelli et al., 2022), LiTS (Bilic et al., 2019), KiTS (Heller et al., 2019), and FLARE (Ma et al., 2024a) have advanced tumor detection and localization. Most benchmarks report only overall averages and miss biases from population and protocol differences. Many are single- or few-site and lack key metadata, which prevents systematic analysis. Recent works (Touchstone (Bassi et al., 2024), OpenMIBOOD (Gutbrod et al., 2025)) reveal generalization gaps but still do not assess group-wise reliability. In this paper, BenchX adds structured annotations (age, sex, race, CT phase, scanner type, voxel spacing) and enables large-scale, comprehensive evaluation across demographic and protocol factors. This provides a fuller view of robustness and bias than accuracy-only reports and reveals specific weaknesses in current models.

3 Dataset

3.1 Dataset Construction

Overview. BenchX aggregates 101,822 abdominal CT scans from six clinical cohorts (Table 1). For each scan, we curate tumor annotations and a unified set of demographic and acquisition metadata to enable subgroup-aware evaluation, including age, sex, race, tumor size, contrast phase, scanner type, voxel spacing, tumor attributes.

Preprocessing. During dataset construction, all scans are converted and stored in NIfTI format for consistency. We convert DICOM intensities to Hounsfield Units (HU) and clip the intensity values to a fixed range of [-1000, 1000]. All scans are preserved at their original spatial resolution.

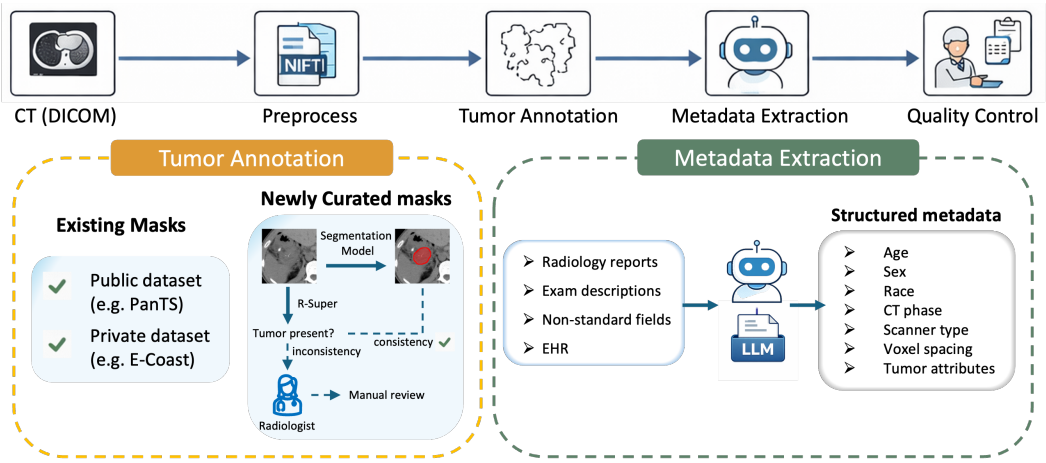


Figure 3: **The BenchX Curation Pipeline.** We unify 101,822 heterogeneous scans into a standardized benchmark using a scalable, semi-automated pipeline. Tumor Annotation: A consistency-check loop where only discordant predictions between segmentation and detection models trigger radiologist review, ensuring expert-level ground truth with minimal manual effort. LLM Metadata Extraction: An automated parser converts unstructured clinical reports into fine-grained attributes (e.g., race, contrast phase), enabling large-scale analysis of model fairness.

Tumor annotations. BenchX uses scan-level tumor presence labels, indicating whether each CT scan contains a pancreatic tumor. For cohorts that provide expert-reviewed tumor masks, such as PanTS (Li et al., 2025) and *E-Coast*, we directly use the provided annotations and convert them to scan-level labels. For cohorts without tumor masks, we create scan-level labels with a double-check process. First, an automatic segmentation model proposes tumor regions with a high-sensitivity setting. Second, an independent detection model (e.g., R-super (Bassi et al., 2025a)) predicts tumor presence at the scan level. Cases with inconsistent results are flagged for manual review. Radiologists then check the original CT scan and reports, and decide the final scan-level label. Annotator roles, clinical background, and additional quality checks are provided in the Appendix A. The definition, collection, and verification of negative cases are detailed in Appendix A.2.

Metadata extraction with LLMs. Clinical cohorts store subgroup and protocol information in multiple sources, including radiology reports, exam descriptions, non-standard metadata fields, and electronic health record data. These sources are often fragmented and inconsistently formatted, which makes large-scale subgroup analysis difficult. To make metadata curation scalable and reproducible, we use a pre-trained LLM to extract structured attributes from free-text reports. We define a fixed schema that specifies the target fields, including age, sex, race, contrast phase, and tumor attributes. The LLM extracts normalized categorical labels according to this schema. It is instructed to quote the exact text spans that support each attribute and to avoid inferring any information that is not explicitly stated in the report. We report extraction accuracy, recall, and inter-rater agreement in Appendix A.3.

Quality control. We apply a two-stage quality control protocol for extracted metadata. First, an LLM performs an automated consistency check by re-extracting key attributes from the original report and comparing them with the normalized outputs. If any extracted attribute does not match the source report or breaks the predefined schema rules, such as more than one phase assigned or unsupported labels, the case is flagged for manual review. A human reviewer then checks the original report and corrects the value if needed, or sets it to `Unknown` when the information is unclear or not explicitly stated. Second, we randomly sample a subset of cases for manual verification and compute attribute level accuracy. We also summarize common error modes, such as ambiguous phase descriptions, missing demographic mentions, and inconsistent terminology. Overall, the observed accuracy and error rates fall within an acceptable range for large scale metadata curation. Detailed statistics and representative examples are provided in the Appendix A.

3.2 Dataset Statistical Analysis

Table 1: **Dataset composition in BenchX.** BenchX comprises **101,822** CT scans from six cohorts. We report scan counts, available sex and age metadata, recorded race coverage, CT phase availability, and cohort proportions.

Dataset	# Scans	Male	Female	Age	Race	Phase‡	Prop. (%)
BenchX	**101,822**	**45,328**	**47,679**	**0–117**	Asian, Black, White, Hispanic, American Indian, Native Hawaiian, Pacific Islander, African, Arab, Balkan, Georgian, Russian, Turkish	**A, V, P, D**	**100.0**
E-Coast	5,102	1,720	2,145	0–107	Asian, Black, White, Hispanic	A, V	5.0
Merlin	25,393	11,100	14,291	1–113	Asian, Black, White, American Indian, Pacific Islander	A, V, P	24.9
PanTS	9,901	2,924	2,358	0–98	Unknown†	A, V, P, D	9.7
N-California	42,096	20,687	21,387	0–117	Asian, Black, White, American Indian, Native Hawaiian	A, V, P	41.3
S-Europe	2,510	833	586	5–91	African, Arab, Balkan, Georgian, Russian, Turkish	A, V, P	2.5
N-Europe	16,820	8,064	6,912	17–102	Unknown†	A, V, P	16.5

† Race metadata is unavailable for these cohorts.
‡ A: arterial; V: venous; P: non-contrast; D: delayed.

We use six external datasets to perform a comprehensive out-of-distribution (OOD) benchmark evaluation. These datasets were collected from hospitals across the globe. Figure 2 compares the metadata available for subgroup evaluation. Compared with MSD-Pancreas and Panorama, BenchX includes race metadata and labeled variation across four CT phases. These factors can be combined to define finer subgroups for evaluation when the corresponding metadata are available. Detailed demographic and lesion statistics are provided in Appendix B. The lesion metadata for BenchX are summarized from available masks, structured annotations, and radiology reports.

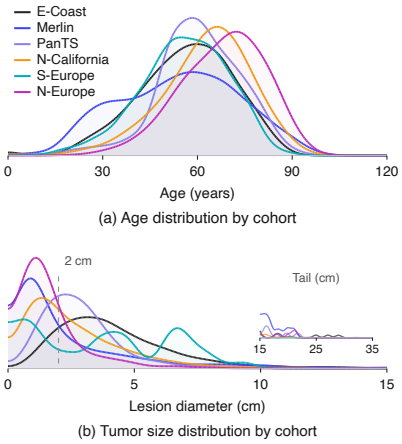


Figure 4: **Dataset statistics of BenchX.** Distributions of (a) age and (b) pancreatic lesion diameter across six cohorts.

Table 1 summarizes BenchX's composition of six cohorts: four cohorts collected by us (*E-Coast*, *N-California*, *S-Europe*, and *N-Europe*) and two public cohorts (Merlin and PanTS). For Merlin, we annotate scan-level tumor presence labels to support tumor detection evaluation. Figure 4 shows the distributions of (a) patient age and (b) pancreatic lesion diameter across these cohorts, providing context for the subgroup analyses below. Representative examples spanning diverse demographics and protocols are shown in Appendix C–H. Data access and release are detailed in Appendix A.4.

4 Benchmark

4.1 Evaluation Protocols

Evaluated models. In BenchX, we evaluated 12 tumor-detection AI models. Appendices I–K provide in-depth descriptions and configuration details for each architecture and framework. We used the same hardware configuration to evaluate all submitted models. Its specifications are: CPU: AMD EPYC 7713 @ 2.0,GHz × 64; GPU: NVIDIA Ampere A100 (80,GB); RAM: 2,TB.

Evaluation Metrics. All evaluated methods are tumor *segmentation* models trained with voxel-wise tumor masks. Our primary evaluation is *scan-level tumor detection*, using scan-level labels that indicate whether each CT scan contains a pancreatic tumor. This scan-level criterion, *if a tumor is present, we detect it*, is widely used in recent works (Xia et al., 2023; Cao et al., 2023) and public benchmarks such as PANORAMA (Alves et al., 2024). Details of the conversion from segmentation

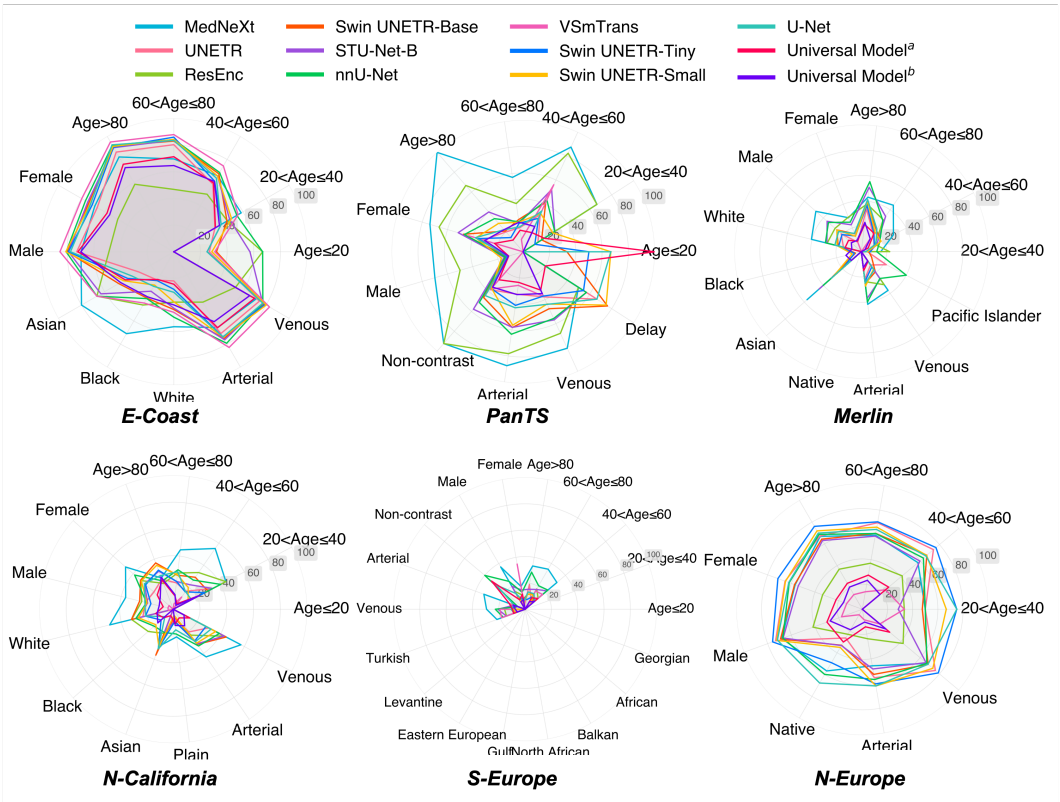


Figure 5: **Benchmark performance biases in BenchX.** Performance varies substantially across cohorts, revealing strong dataset and protocol shifts between institutions. Accuracy is highest on E-Coast, PanTS, and N-Europe but drops on Merlin, N-California, and S-Europe. Arterial and non-contrast scans consistently reduce performance relative to venous CT, and younger and minority populations show lower detection accuracy. Although MedNeXt is generally the most robust model, no architecture performs best across all groups, indicating that average metrics hide important subgroup failures. More results are in Appendix N.

masks to scan-level predictions, including post-processing and threshold settings, are provided in Appendix L.1. We report sensitivity, specificity, AUC, and F1-score for scan-level detection. We report Dice and normalized surface Dice (NSD) for tumor localization.

Evaluation setup. We evaluate 12 tumor-detection models that are trained on a public dataset (Chen et al., 2025). All models are evaluated on BenchX under the same inference setting and hardware configuration, without any additional fine-tuning on BenchX. Importantly, the six cohorts in BenchX are mutually exclusive, with no patient overlap across cohorts.

Data splits. All BenchX data are used only for model evaluation and do not participate in training or fine-tuning. We therefore define two splits, a validation set and a test set. The six cohorts come from different institutions, so there is no patient overlap across cohorts. Within each cohort, we perform a patient-level random split, where all scans from the same patient are kept in the same split. We use a 1:9 ratio for validation and test.

Handling Missing Metadata. Not all patients have complete demographic or protocol information, which is common in real-world clinical practice. For certain attributes such as race or contrast phase, metadata may be unavailable or not explicitly documented. In BenchX, subgroup evaluation is conducted only on cases with definitive labels for the corresponding factor. Cases labeled as *Unknown* are excluded when computing subgroup-specific metrics.

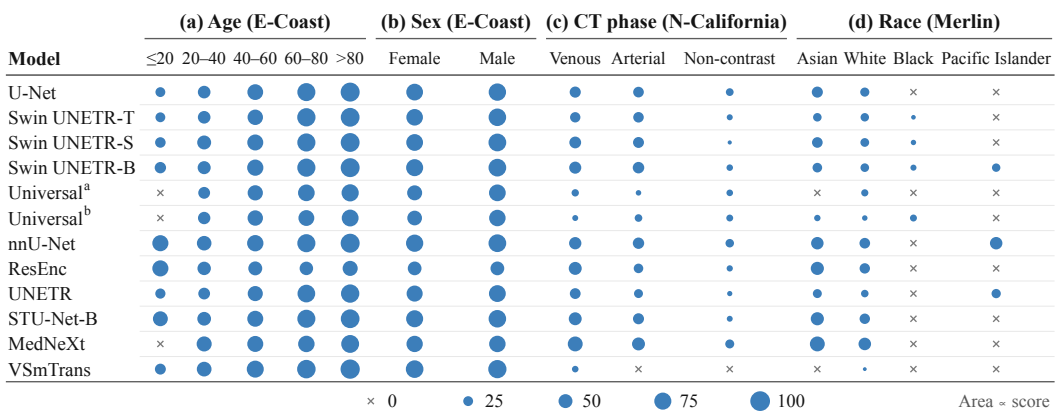


Figure 6: **Subgroup variation in pancreatic tumor detection.** F1 scores (%) of 12 models across (a) age and (b) sex on *E-Coast*, (c) CT phase on *N-California*, and (d) race on *Merlin*. Circle area is proportional to F1 on a shared 0–100 scale. Strong performance in one subgroup does not ensure consistent performance across the others. Full results are provided in Appendix Tables 10, 13, and 12.

4.2 Comprehensive Benchmark Analysis

We evaluate model performance on BenchX across six external cohorts using tumor detection F1 (Figure 5), and report full detection metrics in Appendix N (Tables 10–15). Overall, performance varies widely across cohorts, showing strong site and protocol shifts. MedNeXt is the most stable model overall, but no single model is best for every group.

- **Cohort difficulty varies.** Across models and factors, *Merlin*, *N-California*, and *S-Europe* are consistently more challenging than *E-Coast*, *PanTS*, and *N-Europe*. The best model reaches top-1 subgroup F1 of ∼70–74% on *E-Coast*, *PanTS* and *N-Europe*, but only ∼30–35% on *Merlin*, *N-California* and *S-Europe*, indicating large distribution shifts across cohorts and protocols.
- **Architectural insights.** Aggregating results over all six cohorts and reported factors, MedNeXt achieves the best overall performance. VSmTrans and ResEnc form the next tier and remain competitive across multiple cohorts. This ordering is consistent across all detection metrics (Sen./Spe./AUC/F1).
- **Diversified OOD evaluation is necessary.** For the same model, absolute performance can change by >30% in F1 across cohorts. These large shifts lead to unstable rankings under single-cohort testing, highlighting the need for diversified external evaluation.
- **No model is uniformly strong across factors within a cohort.** Within each cohort, subgroup performance differs across demographic and protocol factors. For example, arterial (and when available non-contrast) phases reduce Sen./F1 by ∼15–30% compared to venous on *E-Coast* and *N-Europe*, even for top-performing models.
- **Consistent hard factors across cohorts.** Across the six cohorts, arterial and non-contrast phases are among the hardest protocol groups, and younger age groups tend to have lower detection accuracy.
- **Error pattern in hard groups.** In many difficult subgroups, specificity remains relatively high while sensitivity drops, suggesting that errors are mainly missed positives. For cohorts with race labels, some minority groups also show pronounced F1 drops.

4.3 Case Study: Demographic and Protocol Variation

Figure 6 examines four factors using available metadata: age and sex on *E-Coast*, CT phase on *N-California*, and race on *Merlin*. Each comparison is made within its corresponding cohort.

(1) Results and analysis.

- **Age.** Nine of 12 models have F1 ≤ 33.3 for patients aged ≤ 20, including zero F1 for Universal^a, Universal^b, and MedNeXt. In contrast, nnU-Net and ResEnc both reach 66.7. Older groups

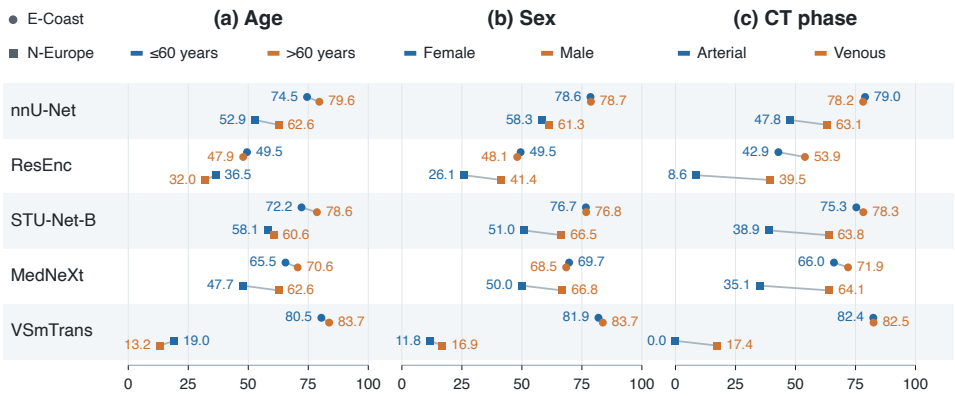


Figure 7: **Model bias with confounder control.** On *E-Coast*, most models show higher F1 in patients aged > 60 years, while sex gaps are small. On *N-Europe*, all models show higher F1 in males and on venous scans, while age patterns vary by model. Subgroup disparities persist after adjustment for the measured factors, with larger sex and CT-phase gaps on *N-Europe*. Additional confounder control experiments are provided in Appendix M.

generally have higher F1, with seven models exceeding 90 for age > 80, although the trend is not monotonic for every model.

- **Sex.** Male F1 exceeds female F1 for all 12 models, but the gap is below 10% for 10 models. Universal^a and Universal^b have larger gaps of 14.1 and 14.2%. VSmTrans leads both groups, with 79.8 for females and 85.5 for males.
- **CT phase.** Venous scans yield the highest F1 for 10 of 12 models. MedNeXt leads all three phases, with 57.1, 43.2, and 20.8 for venous, arterial, and non-contrast scans, respectively. Every model scores at most 20.8 on non-contrast scans, indicating a shared weakness despite differences in architecture.
- **Race.** Eight models have zero F1 for Black patients and nine for Pacific Islander patients. MedNeXt leads the Asian and White groups (57.1 and 40.5) but scores zero in both remaining groups. The best F1 for Black patients is only 12.5 (Universal^b), whereas nnU-Net reaches 40.0 for Pacific Islander patients. Thus, subgroup rankings differ even within the same cohort.

(2) Clinical justification. The following clinical context motivates possible explanations; the subgroup comparisons alone do not establish their causes.

- **Age.** Pancreatic volume increases during childhood and adolescence, remains relatively stable from ages 20 to 60, and declines thereafter; pancreatic fat also changes with age (Saisho et al., 2007). These changes may alter the image features available to a detector, but organ size alone cannot explain the observed advantage in older patients.
- **Sex.** Pancreatic volume and fat distribution differ on average between sexes (Saisho et al., 2007), providing possible sources of variation in CT appearance. Whether they explain the F1 gap requires accounting for body size, lesion characteristics, and subgroup disease prevalence.
- **CT phase.** Contrast timing changes tumor–pancreas contrast; dedicated pancreatic-phase imaging can provide greater contrast than later hepatic-phase imaging (Lu et al., 1996). The venous advantage observed here therefore describes these models on this cohort, rather than a universal clinical ordering of phases. Protocol representation during training and differences in case mix are plausible additional contributors.
- **Race.** Clinical populations may differ in age, disease prevalence, and acquisition conditions. Prior imaging audits account for subgroup composition when evaluating disparities (Glocker et al., 2023). The gaps on *Merlin* warrant such analysis, with subgroup counts and uncertainty estimates, before attributing them to training imbalance or a biological mechanism.

4.4 Confounder Control Analysis

Demographic factors, CT protocols, and lesion characteristics can vary together, making unadjusted subgroup gaps difficult to interpret. We therefore assess whether these gaps persist after balancing the

other measured factors. We evaluate five models on *E-Coast* and *N-Europe* using scans with complete metadata and available predictions. For each age, sex, or CT-phase comparison, we jointly standardize the remaining measured factors, lesion status, and lesion size within each cohort. Weighting details and 95% confidence intervals are provided in Appendix M.3; sensitivity results are in Appendix M.1.

Results. Figure 7 shows subgroup F1 after confounder control. On *E-Coast*, patients aged > 60 have higher F1 for most models, while sex gaps are below 1.9%. On *N-Europe*, all models have higher F1 on venous scans, with gaps of 15.3–30.9%; sex gaps range from 3.0 to 16.8% in favor of males. Subgroup disparities persist after balancing the measured factors, suggesting that the covariates considered here do not fully account for the observed gaps. The larger sex and CT-phase gaps on *N-Europe* further indicate that subgroup robustness is cohort-dependent. These findings highlight the need to evaluate demographic and protocol biases across diverse clinical settings and to develop mitigation strategies for subgroups with lower performance.

4.5 Synthetic Data for Domain Adaptation

We hypothesize that synthetic data can help when some demographic groups or scan settings have few training cases. This can reduce subgroup gaps by adding more diverse examples. We test this hypothesis on the *E-Coast* dataset.

We modified existing conditional diffusion model approaches (Wu et al., 2024; Chen et al., 2024; Li et al., 2026) to generate synthetic tumors conditioned on patient demographics and imaging protocols. We call this method SynthX. Specifically, we adapt existing conditional diffusion models to generate synthetic tumors by conditioning on patient sex, age, and image contrast phases (non-contrast, arterial, venous). This allows us to synthesize tumors with specific demographic and protocol attributes, filling gaps in underrepresented subgroups. Implementation details are provided in the Appendix K. We compare the MedNeXt baseline with the SynthX setting on *E-Coast*.

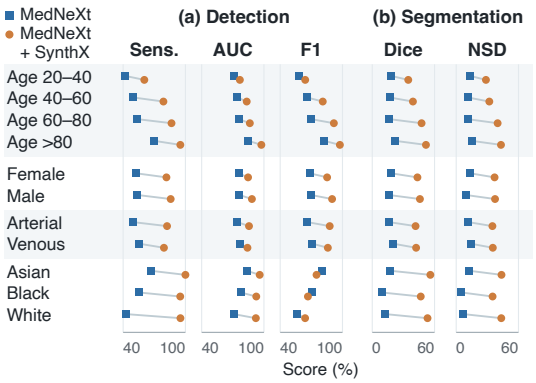


Figure 8: **Detection and segmentation with SynthX.** SynthX achieves higher sensitivity, Dice, and NSD across all shown *E-Coast* subgroups. These gains across demographic and CT-phase groups support SynthX as a promising solution for addressing demographic and protocol biases.

Results. SynthX achieves higher sensitivity, Dice, and NSD across all reported *E-Coast* subgroups (Figure 8). These gains support SynthX as a promising solution for addressing demographic and protocol biases.

5 Conclusion

BenchX is a global benchmark built from 101,822 CT scans representing different ages, races, and imaging protocols. Our evaluation exposes critical limitations in current tumor detection models that are often overlooked: **(1) Age dependent performance degradation is severe.** Younger patients (Age ≤ 40) show dramatic performance drops with near-zero sensitivity. Top-performing models completely fail on patients under 20 years old, a vulnerability rarely discussed in medical imaging literature. **(2) Imaging protocol is a dominant factor, rivaling architecture.** Venous phase scans yield F1 scores up to 57.1%, while non-contrast scans produce near-zero sensitivity. Single-protocol benchmarks fundamentally misrepresent real-world performance. **(3) Racial disparities are severe and systemic.** Black and Pacific Islander patients exhibit near-zero sensitivity across most top-ranked models, suggesting limited model generalization across patient populations. **(4) Standard fairness metrics mask catastrophic failures.** Overall performance can appear acceptable (70–80% F1) while demographic or protocol subgroups collapse to zero sensitivity. Average metrics alone are insufficient for safety-critical medical applications. **(5) Synthetic-data results vary by metric.** The SynthX setting has higher sensitivity, Dice, and NSD across the reported *E-Coast* subgroups, but lower F1 for Asian and Black patients.

AI USE STATEMENT

We used ChatGPT to assist with manuscript refinement and visualization code. The authors take full responsibility for the methodology, results, citations, and final manuscript.

ETHICS STATEMENT

BenchX combines public datasets and institutional clinical cohorts. The study received institutional review board (IRB) approval. Private data are de-identified and used under institutional governance. Institution-identifying IRB details are withheld to preserve double-blind review and will be provided in the final version.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will release the code, trained models, and shareable European cohorts (*S-Europe* and *N-Europe*) after de-identification. For Merlin and PanTS, we will provide scripts that convert the source datasets into an evaluation-ready format, together with shareable derived labels and metadata. Following this release, 54,624 CT scans across these four cohorts will be publicly accessible, supporting direct evaluation of demographic and protocol biases under the settings used in this paper. Access-restricted cohorts, including *N-California*, will support third-party evaluation through a challenge. Public source datasets remain subject to their original access conditions. Cohort-specific release contents and access arrangements are provided in Appendix A.4.

REFERENCES

N Alves, M Schuurmans, D Rutkowski, et al. The panorama study protocol: Pancreatic cancer diagnosis-radiologists meet ai. zenodo, 2024.

Natalia Alves, Megan Schuurmans, Dawid Rutkowski, Anindo Saha, Pierpaolo Vendittelli, Nancy Obuchowski, Marjolein H Liedenbaum, Ingfrid S Haldorsen, Anders Molven, Derya Yakar, et al. Artificial intelligence and radiologists in pancreatic cancer detection using standard of care ct scans (panorama): an international, paired, non-inferiority, confirmatory, observational study. *The Lancet Oncology*, 27(1):116–124, 2026.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):1–13, 2022.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Pedro RAS Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Yannick Kirchhoff, Maximilian Rokuss, Ziyan Huang, Jin Ye, Junjun He, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus H. Maier-Hein, Paul Jaeger, Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Yong Xia, Zhaohu Xing, Lei Zhu, Yousef Sadegheih, Afshin Bozorgpour, Pratibha Kumari, Reza Azad, Dorit Merhof, Pengcheng Shi, Ting Ma, Yuxin Du, Fan Bai, Tiejun Huang, Bo Zhao, Haonan Wang, Xiaomeng Li, Hanxue Gu, Haoyu Dong, Jichen Yang, Maciej A. Mazurowski, Saumya Gupta, Linshan Wu, Jiaxin Zhuang, Hao Chen, Holger Roth, Daguang Xu, Matthew B. Blaschko, Sergio Decherchi, Andrea Cavalli, Alan L. Yuille, and Zongwei Zhou. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 37:15184–15201, 2024. URL https://github.com/MrGiovanni/Touchstone.

Pedro RAS Bassi, Wenxuan Li, Jieneng Chen, Zheren Zhu, Tianyu Lin, Sergio Decherchi, Andrea Cavalli, Kang Wang, Yang Yang, Alan L Yuille, and Zongwei Zhou. Learning segmentation from radiology reports. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 305–315. Springer, 2025a. URL https://github.com/MrGiovanni/R-Super.

Pedro RAS Bassi, Xinze Zhou, Wenxuan Li, Szymon Płotka, Jieneng Chen, Qi Chen, Zheren Zhu, Jakub Przado, Ibrahim E Hamamci, Sezgin Er, Xiaoxi Chen, Mehmet Can Yavuz, Yu-Cheng Chou, Tianyu Lin, Kang Wang, Yucheng Tang, Jaroslaw B Cwikla, Sergio Decherchi, Andrea Cavalli, Yang Yang, Alan L Yuille, and Zongwei Zhou. Scaling artificial intelligence for multi-tumor early detection with more reports, fewer masks. *arXiv preprint arXiv:2510.14803*, 2025b. URL https://github.com/MrGiovanni/R-Super.

Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019.

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Robbie Holland, Cesar Truyts, Christian Bluethgen, Yufu Wu, Long Lian, Malte Engmann Kjeldskov Jensen, Sophie Ostmeier, Maya Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson Amaro Junior, Neera Ahuja, Jason Fries, Nigam H. Shah, Greg Zaharchuk, Marc Willis, Adam Yala, Andrew Johnston, Robert D. Boutin, Andrew Wentland, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, and Akshay S. Chaudhari. Merlin: a computed tomography visionlanguage foundation model and dataset. *Nature*, 652(8112):13181328, March 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10181-8. URL http://dx.doi.org/10.1038/s41586-026-10181-8.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature medicine*, 29(12):3033–3043, 2023.

M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

Jieneng Chen, Yingda Xia, Jiawen Yao, Ke Yan, Jianpeng Zhang, Le Lu, Fakai Wang, Bo Zhou, Mingyan Qiu, Qihang Yu, Mingze Yuan, Wei Fang, Yuxing Tang, Minfeng Xu, Jian Zhou, Yuqian Zhao, Qifeng Wang, Xianghua Ye, Xiaoli Yin, Yu Shi, Xin Chen, Jingren Zhou, Alan Yuille, Zaiyi Liu, and Ling Zhang. Cancerunit: Towards a single unified model for effective detection, segmentation, and diagnosis of eight major cancers using a large collection of ct scans. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21270–21281, 2023. doi: 10.1109/ICCV51070.2023.01950.

Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11147–11158, 2024. URL https://github.com/MrGiovanni/DiffTumor.

Qi Chen, Xinze Zhou, Chen Liu, Hao Chen, Wenxuan Li, Zekun Jiang, Ziyan Huang, Yuxuan Zhao, Dexin Yu, Junjun He, Yefeng Zheng, Ling Shao, Alan Yuille, and Zongwei Zhou. Scaling tumor segmentation: Best lessons from real and synthetic data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 24001–24013, 2025. URL https://github.com/BodyMaps/AbdomenAtlas2.0.

Moon S Chen Jr, Primo N Lara, Julie HT Dang, Debora A Paterniti, and Karen Kelly. Twenty years post-nih revitalization act: enhancing minority participation in clinical trials (empact): laying the groundwork for improving minority clinical trial accrual: renewing the case for enhancing minority participation in cancer clinical trials. *Cancer*, 120:1091–1096, 2014.

Ben Glocker, Charles Jones, Mélanie Roschewitz, and Stefan Winzeck. Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence*, 5(6):e230060, 2023.

Max Gutbrod, David Rauber, Danilo Weber Nunes, and Christoph Palm. Openmibood: Open medical imaging benchmarks for out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25874–25886, 2025.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584, 2022.

Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, 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.

Anita Holdcroft. Gender bias in research: how does it affect evidence based medicine? *Journal of the Royal Society of Medicine*, 100(1):2–3, 2007. doi: 10.1177/014107680710000102. URL https://doi.org/10.1177/014107680710000102. PMID: 17197669.

Can Hu, Yingda Xia, Zhilin Zheng, Mengxuan Cao, Guoliang Zheng, Shangqi Chen, Jiancheng Sun, Wujie Chen, Qi Zheng, Siwei Pan, Yanqiang Zhang, Jiahui Chen, Pengfei Yu, Jingli Xu,

Jianwei Xu, Zhongwei Qiu, Tiancheng Lin, Boxiang Yun, Jiawen Yao, Wenchao Guo, Chen Gao, Xianghui Kong, Keda Chen, Zhengle Wen, Guanxia Zhu, Jinfang Qiao, Yibo Pan, Huan Li, Xijun Gong, Zaisheng Ye, Weiqun Ao, Lei Zhang, Xing Yan, Yahan Tong, Xinxin Yang, Xiaozhong Zheng, Shufeng Fan, Jielu Cao, Cheng Yan, Kangjie Xie, Shengjie Zhang, Yao Wang, Lin Zheng, Yingjie Wu, Zufeng Ge, Xiyuan Tian, Xin Zhang, Yan Wang, Ruolan Zhang, Yizhou Wei, Weiwei Zhu, Jianfeng Zhang, Hanjun Qiu, Miaoguang Su, Lei Shi, Zhiyuan Xu, Ling Zhang, and Xiangdong Cheng. Ai-based large-scale screening of gastric cancer from noncontrast ct imaging. *Nature Medicine*, 31(9):3011–3019, 2025. doi: 10.1038/s41591-025-03785-6. URL https://doi.org/10.1038/s41591-025-03785-6.

Kai Hu, Jinhao Li, Yuan Zhang, Xiongjun Ye, and Xieping Gao. One-to-multiple: A progressive style transfer unsupervised domain-adaptive framework for kidney tumor segmentation. *Advances in Neural Information Processing Systems*, 37:24496–24522, 2024.

Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, et al. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv preprint arXiv:2304.06716*, 2023.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. *arXiv preprint arXiv:2404.09556*, 2024.

Rahul Kumar Jain, Takahiro Sato, Taro Watasue, Tomohiro Nakagawa, Yutaro Iwamoto, Xianhua Han, Lanfen Lin, Hongjie Hu, Xiang Ruan, and Yen-Wei Chen. Unsupervised domain adaptation using adversarial learning and maximum square loss for liver tumors detection in multi-phase ct images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1536–1539. IEEE, 2022.

Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30(4): 1166–1173, 2024.

Wenxuan Li, Xinze Zhou, Qi Chen, Tianyu Lin, Pedro RAS Bassi, Xiaoxi Chen, Chen Ye, Zheren Zhu, Kai Ding, Heng Li, Kang Wang, Yang Yang, Yucheng Tang, Daguang Xu, Alan L Yuille, and Zongwei Zhou. Pants: The pancreatic tumor segmentation dataset. In *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025. URL https://github.com/MrGiovanni/PanTS.

Xinran Li, Yi Shuai, Chen Liu, Qi Chen, Tianyu Lin, Pengfei Guo, Dong Yang, Can Zhao, Pedro RAS Bassi, Daguang Xu, Kang Wang, Yang Yang, Alan L Yuille, and Zongwei Zhou. Text-driven tumor synthesis. *IEEE Transactions on Medical Imaging*, 2026. URL https://github.com/MrGiovanni/TextoMorph.

Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. 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. URL https://github.com/ljwztc/CLIP-Driven-Universal-Model.

Tiange Liu, Qingze Bai, Drew A Torigian, Yubing Tong, and Jayaram K Udupa. Vsmtrans: A hybrid paradigm integrating self-attention and convolution for 3d medical image segmentation. *Medical image analysis*, 98:103295, 2024.

William Lotter. Acquisition parameters influence ai recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias. *Nature communications*, 15(1):7465, 2024.

D. S. Lu, S. Vedantham, R. M. Krasny, B. Kadell, W. L. Berger, and H. A. Reber. Two-phase helical CT for pancreatic tumors: pancreatic versus hepatic phase enhancement of tumor, pancreas, and vascular structures. *Radiology*, 199(3):697–701, 1996. doi: 10.1148/radiology.199.3.8637990.

Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Ershuai Wang, Qin Zhou, Ziyan Huang, Pengju Lyu, Jian He, and Bo Wang. Automatic organ and pan-cancer segmentation in abdomen ct: the flare 2023 challenge, 2024a. URL https://arxiv.org/abs/2408.12534.

Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Ershuai Wang, Qin Zhou, Ziyan Huang, Pengju Lyu, Jian He, and Bo Wang. Automatic organ and pan-cancer segmentation in abdomen ct: the flare 2023 challenge. *arXiv preprint arXiv:2408.12534*, 2024b.

Institute of Medicine. *Women and Health Research: Ethical and Legal Issues of Including Women in Clinical Studies, Volume 1*. The National Academies Press, Washington, DC, 1994. ISBN 978-0-309-04992-4. doi: 10.17226/2304. URL https://nap.nationalacademies.org/catalog/2304/women-and-health-research-ethical-and-legal-issues-of-including.

P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 405–415. Springer, 2023.

Y. Saisho, A. E. Butler, J. J. Meier, T. Monchamp, M. Allen-Auerbach, R. A. Rizza, and P. C. Butler. Pancreas volumes in humans from birth to age one hundred taking into account sex, obesity, and presence of type-2 diabetes. *Clinical Anatomy*, 20(8):933–942, 2007. doi: 10.1002/ca.20543.

Lawrence Sirovich and Michael Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4(3):519–524, 1987.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. 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.

Yu Tian, Min Shi, Yan Luo, Ava Kouhana, Tobias Elze, and Mengyu Wang. Fairseg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. *arXiv preprint arXiv:2311.02189*, 2023.

M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586–591, 1991. doi: 10.1109/CVPR.1991.139758.

Ryan Wang, Po-Chih Kuo, Li-Ching Chen, Kenneth Patrick Seastedt, Judy Wawira Gichoya, and Leo Anthony Celi. Drop the shortcuts: image augmentation improves fairness and decreases ai detection of race and other demographics from medical images. *EBioMedicine*, 102, 2024.

Linshan Wu, Jiaxin Zhuang, Xuefeng Ni, and Hao Chen. Freetumor: Advance tumor segmentation via large-scale tumor synthesis. *arXiv preprint arXiv:2406.01264*, 2024.

Yingda Xia, Qihang Yu, Linda Chu, Satomi Kawamoto, Seyoun Park, Fengze Liu, Jieneng Chen, Zhuotun Zhu, Bowen Li, Zongwei Zhou, Yongyi Lu, Yan Wang, Wei Shen, Lingxi Xie, Yuyin Zhou, Christopher Wolfgang, Ammar Javed, Daniel Fadaei Fouladi, Shahab Shayesteh, Jefferson Graves, Alejandra Blanco, Eva S. Zinreich, Miriam Klauss, Philipp Mayer, Benedict Kinny-Köster, Kenneth Kinzler, Ralph H. Hruban, Bert Vogelstein, Alan L. Yuille, and Elliot K. Fishman. The

felix project: Deep networks to detect pancreatic neoplasms. *medRxiv*, 2023. doi: 10.1101/2022.09.24.22280071. URL https://www.medrxiv.org/content/early/2023/02/27/2022.09.24.22280071.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501, 2018.

Yuzhe Yang, Haoran Zhang, Judy W. Gichoya, Dina Katabi, Marzyeh Ghassemi, et al. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, 30:28382848, 2024. doi: 10.1038/s41591-024-03113-4.

Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 508–518. Springer, 2023.

Yuwei Yin, Zhixian Tang, Zheng Huang, Mingxuan Wang, and Huachun Weng. Source free domain adaptation for kidney and tumor image segmentation with wavelet style mining. *Scientific Reports*, 14(1):24849, 2024.