

Interactive Lesion Segmentation in Whole-Body PET/CT and Longitudinal CT: The autoPET/CT-IV Challenge

Pauline Ornela Megne Choudja

Pauline.Megne@med.uni-tuebingen.de

University Hospital of Tübingen

Jakob Dextl

LMU University Hospital

Zdravko Marinov

Karlsruhe Institute of Technology

Niels Rocholl

Radboud University Medical Center

Temke Kohlbrandt

Fraunhofer Institute for Digital Medicine MEVIS

Antoine Sanner

University Medical Center Mainz <https://orcid.org/0000-0002-4917-9529>

Katharina Jeblick

LMU University Hospital

Balthasar Schachtner

LMU University Hospital

Andreas Mittermeier

LMU University Hospital

Anna Theresa Stüber

LMU University Hospital

Matthias Fabritius

Walter Brendel Centre of Experimental Medicine

Junwei Huang

School of Biomedical Engineering, Tsinghua University

Yingqi Hao

Tsinghua University

Lei Xiang

Subtle Medical Inc.

Qiyuan Tian

Tsinghua University <https://orcid.org/0000-0002-8350-5295>

Yannick Kirchhoff

German Cancer Research Center (DKFZ)

Maximilian Rokuss

German Cancer Research Center (DKFZ)

Fabian Isensee

German Cancer Research Center <https://orcid.org/0000-0002-3519-5886>

Klaus Maier-Hein

German Cancer Research Center

Zacharia Mesbah

Rouen Normandy University

Solène Perret

Rouen Normandy University

Léo Mottay

Rouen Normandy University

Yang Xing

University of Florida

Kuang Gong

University of Florida <https://orcid.org/0000-0002-2669-2610>

Moona Mazher

University College London

Abdul Qayyum

National Heart and Lung Institute, Faculty of Medicine, Imperial College London

Lalith Kumar Shiyam Sundar

LMU University Hospital

Tanja Loßau

Fraunhofer Institute for Digital Medicine MEVIS

Jan Moltz

Fraunhofer Institute for Digital Medicine MEVIS

Alessa Hering

Radboudumc <https://orcid.org/0000-0002-7602-803X>

Dirk Graafen

University Medical Center Mainz

Ahmed Othman

University Medical Center of the Johannes Gutenberg University Mainz

Andreas Wagner-Defant

University Hospital of Tübingen

Felix Peisen

University Hospital of Tübingen

Sergios Gatidis

Stanford University

Adrien Holzgreve

Dept. of Nuclear Medicine, University Hospital of Munich, LMU Munich, Munich, Germany

<https://orcid.org/0000-0002-9566-8438>

Konrad Klimek

LMU University Hospital

Rudolf Werner

LMU Hospital

Clemens C. Cyran

LMU University Hospital

Michael Ingrisich

LMU University Hospital, LMU Medizin, Ludwig-Maximilians-Universität München, Munich, Germany

<https://orcid.org/0000-0003-0268-9078>

Thomas Kuestner

University Hospital Tübingen <https://orcid.org/0000-0002-0353-4898>

Article

Keywords:

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10893740/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: **Yes** there is potential Competing Interest. We declare the following conflicts of interest: Rudolf A. Werner has received speaker honoraria from Novartis. Katharina Jeblick is one of the founders and CEOs of Olatu Medical Technologies GmbH. The work is not linked to the company's activities.

Abstract

Deep learning enables automated tumour segmentation in whole-body positron emission tomography and computed tomography, but robustness remains limited by heterogeneous tracer uptake, anatomical variability, and inter-center differences. We report the fourth autoPET challenge, which evaluated interactive lesion segmentation using incremental simulated user prompts and longitudinal lesion segmentation from paired baseline and follow-up scans. In the interactive task (Task 1, ~ 30,000 training and ~ 3,000 test lesions), submitted methods improved with additional user input and achieved a mean Dice similarity coefficient of 0.74 at the final interaction step, with performance remaining robust across tracers and centers. Longitudinal segmentation (Task 2, ~ 7,500 training and ~ 4,500 test lesions) was substantially more challenging, with a best mean Dice similarity coefficient of 0.38. Here, we show that interactive segmentation can combine automated inference with iterative expert input while maintaining performance across heterogeneous imaging domains, whereas longitudinal lesion prediction remains limited by temporal and anatomical variability.

Introduction

Cancer imaging frequently presents clinicians with two fundamentally different questions: what is the current disease burden, and how is it changing over time?. While these questions are central to clinical decision-making, answering them efficiently requires both accurate segmentation of existing lesions and careful tracking of their evolution. In clinical practice, radiologists and nuclear medicine physicians rarely work in isolation: initial lesion outlines are often generated using semi-automatic intensity-based methods, such as thresholds based on standardized uptake values (SUV¹) or Hounsfield Units (HU), and then refined through expert review^{2,3}, while follow-up scans are compared against baseline images to assess progression or response to therapy.

Despite these thresholding-based starting points, manual lesion segmentation remains labor-intensive and subject to inter-reader variability, with increasing imaging volumes and case complexity further compounding the burden⁴⁻⁷. These constraints motivated the launch of the autoPET challenge, initiated to stimulate the development of automated, Deep Learning (DL)-based approaches for tumor lesion segmentation. Across its first three editions, the autoPET challenge progressively expanded the scope and difficulty of the task. The first edition⁸ established the feasibility of large-scale, fully automated whole-body PET/CT lesion segmentation, providing a common benchmark and demonstrating the promise of DL-based solutions. The second⁹ and third¹⁰ editions extended this foundation by probing generalization across institutions, scanners, and tracers, exposing the sensitivity of models to domain shift and heterogeneous imaging conditions. Together, these editions created a robust methodological and data-driven framework for automated lesion segmentation, while also revealing persistent bottlenecks. Even the best-performing models continued to struggle with small lesions and remained domain dependent. More importantly, purely automated pipelines did not fully reflect clinical reality, where expert oversight, iterative correction, and longitudinal comparison are integral to decision-making.

These observations indicated that achieving reliable and clinically meaningful lesion segmentation may require moving beyond better automation toward workflows that explicitly integrate human expertise and temporal reasoning.

In response to these limitations, the fourth edition of the autoPET/CT challenge shifts the focus from fully automated segmentation towards interactive and hybrid human-AI workflows through two complementary tasks. Task 1 focuses on human-in-the-loop segmentation of baseline scans, where models provide initial lesion masks that can be interactively reviewed and corrected by an expert. By integrating algorithmic proposals with simulated human refinement, Task 1 mirrors real-world workflows in which automated tools augment expert assessment. Task 2 addresses the temporal dimension of oncological imaging by requiring models to predict lesions in follow-up scans based on baseline data. This transforms segmentation into a longitudinal prediction task: algorithms must anticipate lesion emergence, regression, and spatial redistribution. Accurately forecasting these changes demands models to capture disease dynamics and to maintain lesion-level consistency across timepoints, providing insights into tumor progression that go beyond static delineation. Task 1 used the established autoPET training data^{11,12}, whereas Task 2 introduced a newly released longitudinal CT dataset¹³.

Here, we present the results of the autoPET/CT-IV challenge: Automated Lesion Segmentation in Whole-Body PET/CT and Longitudinal CT - The human frontier, hosted at the Medical Imaging Computing and Computed Assisted Intervention Society (MICCAI) 2025. We describe the challenge structure, dataset composition, evaluation metrics, and participating methods, and analyze the resulting performance. In particular, we examine the effect of user-provided clicks on segmentation accuracy and efficiency, assess ranking stability, and investigate the impact of test-data variability on model performance. We characterize the capabilities and limitations of interactive approaches for lesion segmentation in PET/CT and longitudinal CT imaging, and provide insight into how human-AI collaboration can be designed to support robust, clinically deployable segmentation tools.

Results

Challenge submission and selected participating teams

A total of 191 participants from Asia, Europe, the Americas, and Africa registered for the challenge, forming 14 teams across both tasks. In the preliminary phase, five teams submitted algorithms for Task 1, and three for Task 2. In the final phase, seven teams submitted algorithms (five for Task 1 and two for Task 2), and the three best-performing teams, in accordance with the criteria, were designated as challenge winners.

Six of the seven submitted algorithms used a 3D U-Net¹⁴ architecture, implemented within the nnU-Net framework¹⁵, whereas one team applied a multistage strategy combining a masked autoencoder¹⁶ with a BiXLSTM¹⁷-ResNet encoder. Additional algorithm details, including baselines, are provided in the appendix (pp 22–24).

Task 1: Single-staging whole-body PET/CT

Performance metrics for the five submitted algorithms, alongside the two baseline models and the ensemble model, ranked according to the challenge scoring scheme, are presented in Fig. 1. The best-performing team (BIRTH) achieved a mean DSC@last of 0.74, with FNV@last of 2.17 ml, and FPV@last of 1.30 ml. Across the top four ranked submissions, DSC@last ranged from 0.71 to 0.74, FNV@last from 1.75 to 4.22 ml, and FPV@last from 1.29 to 2.66 ml. AUC-based measures reflected progressive accuracy gains with incremental user interaction: AUC-DSC ranged from 6.73 to 7.24, AUC-FNV from 25.66 to 64.05 ml, and AUC-FPV from 13.74 to 26.74 ml.

Stratified analysis by tracers and acquisition centers (appendix p 2) showed robust performance of the two top-ranked models across domains. Higher mean DSC@last was observed for UKT cohort (FDG-UKT: 0.80–0.82; PSMA-UKT: 0.67–0.70) compared with LMU cohort (FDG-LMU: 0.70, PSMA-LMU: 0.72–0.74). Mean FNV@last was lower for UKT (0.37–0.47 ml) than for LMU (0.84–6.65 ml), whereas FPV@last was reduced in LMU (0.39–1.44 ml) compared to UKT (0.99–3.96 ml). These patterns were consistent across AUC-derived metrics. Notably, the submitted methods outperformed, on average, those reported in autoPET III results¹⁰ (Fig. 2). In particular, the best-performing autoPET/CT IV submission surpassed the top algorithm reported in autoPET III, with consistent improvements in DSC ranging from 5.9% (FDG-LMU) to 21.9% (PSMA-UKT) and substantially reduced FPV (up to -77.1% in FDG-LMU), suggesting a meaningful gain in generalizability.

Among the two provided baselines, nnU-Net model ranked fifth with a competitive mean DSC@last of 0.63 and FNV@last of 3.00 ml, but limited by a substantially high FPV@last (10.0 ml). The Sliding-Window FASTEDIT (SW-FASTEDIT) baseline was outperformed by all submitted models. A comprehensive overview of all computed metrics across centers and tracers for each submitted algorithm is provided in appendix (pp 3–4).

Figure 1: Task 1. Performance distribution of all ranked algorithms across the six evaluation metrics: (A) Dice Similarity Coefficient (DSC), (B) False Positive Volume (FPV), and (C) False Negative Volume (FNV) at the last interactive step (i.e. 10 foreground and 10 background clicks), and (D-F) the respective area-under-the-curve (AUC). Bars are ordered from left to right according to the official challenge ranking. The ensemble, nnU-Net, and SW-FASTEDIT baselines are out of competition (†). The star symbolizes the average performance.

The out-of-competition ensemble integrated predictions from all submitted algorithms and both baselines. It achieved the strongest overall performance with mean DSC@last 0.75, FNV@last 1.18 ml, and FPV@last 1.54 ml. Bootstrap analysis with 1000 resamples confirmed ranking robustness (appendix p 5), indicating that some methods consistently outperformed others across varied data subsets.

Segmentation quality as a function of the number of clicks (NoC) across tracers and centers for both the baseline and ensemble models is shown in appendix p 6, providing a comparison between the initial state at challenge start and the best achievable performance at challenge completion, respectively.

Progressive improvements with increased user interaction were observed across models: DSC@last and AUC-DSC increased with additional clicks, while FNV@last and AUC-FNV decreased, indicating refinement and reduction of missed lesions. By contrast, FPV@last and AUC-FPV showed an increase with more clicks, suggesting some degree of FP propagation during refinement, especially for the PSMA-UKT subgroup. Analysis of these FPV patterns (appendix p 7) showed that, across datasets, increasing FPV categories were generally associated with higher lesion burden, reflected by both an increased number of lesions and larger lesion volumes, consistent with Spearman correlation (appendix p 8). However, looking into each subgroup revealed that the PSMA-UKT subgroup, despite exhibiting the highest FPV values, was associated with comparatively smaller lesion burden than the other subgroups. In addition, cases with non-zero FPV tended to include lesions with more heterogeneous uptake values (SUV max or SUV mean). By contrast, smaller lesions and lesions with very low tracer uptake were less frequently detected than larger lesions and lesions with higher uptake (appendix p 9). Detection performance was generally consistent across subgroups, with a mean sensitivity exceeding 0.6, including in the low-uptake setting, with the exception of PSMA-UKT.

Figure 2: Task 1. Comparison of all submissions from autoPET/CT-IV (IV) to the models from autoPET III (III). Percentage improvement (from III to IV) is displayed between paired bars. Overall, autoPET IV algorithms demonstrate superior segmentation quality with reduced FPV (except in PSMA-UKT) and FNV across all cohorts, indicating enhanced generalizability. Error bars represent the 95% confidence interval (CI).

Performance summaries at 1, 5, and 10 clicks (DSC@1/5/10; FNV@1/5/10; FPV@1/5/10) with the number of clicks required to achieve DSC thresholds of 0.85 and 0.90 (denoted as NoC@85 and NoC@90) are provided in appendix p 10. Overall, highest DSC and lowest FNV were typically achieved at 10 clicks, whereas FPV was lowest at 5 clicks. Across methods, ~ 9 clicks were required on average to reach $DSC \geq 0.85$.

Visual inspection of common error maps revealed spatial patterns of disagreement across algorithms (Fig. 3). Persistent errors occurred in areas of low contrast (Fig. 3a), and in clusters of small metastatic lesions, particularly in the spine (Fig. 3b, c). Some errors remained despite increasing user interaction, whereas others were reduced (appendix p 11), indicating heterogeneous sensitivity of error regions to interactive refinement.

Figure 3: Task 1. Qualitative analysis of common segmentation error maps at different levels of user interaction. Segmentation errors from each algorithm are overlaid on PET images for increasing numbers of foreground and background clicks (0, 1, 5, and 10). False positive errors (FP) are shown in red and false negative (FN) errors in blue. Colour intensity reflects the number of individual algorithms (excluding the Ensemble model) that misclassify a given voxel, ranging from 0 (all algorithms correct at that voxel) to 7 (seven algorithms incorrect).

Task 2: Longitudinal whole-body CT

Performance of the two submitted algorithms and the baseline is shown in Fig. 4. Metrics are depicted for a lesion click in FU timepoint to infer longitudinal segmentation performance. The best-performing team (LesionLocator) achieved a mean DSC@last = 0.38, with FNV@last = 1.09 ml, and FPV@last = 0.63 ml. Team nRocholl ranked second with a mean DSC@last = 0.30, FNV@last = 2.55 ml, FPV@last = 1.48 ml, while the baseline achieved a mean DSC@last = 0.25, FNV@last = 1.60 ml, and FPV@last = 0.97 ml. The submission from team nRocholl was considered out of competition due to institutional involvement as co-organizers.

Figure 4: Task 2. Lesion-level segmentation performance at the follow-up (FU) time point for all ranked algorithms across the three evaluation metrics (A, B, C). Bars are ordered from left to right according to the official challenge ranking. The nRocholl team model and LongiSeg baseline are out of competition (†). The star symbolizes the average performance.

Center-level analysis (appendix p 12) showed stable performance of the winning method across both centers. In contrast, the other two algorithms performed better on the out-of-domain UM cohort, which contains more cases ($n = 195$) and lower overall lesion burden (total volume $\approx 6.15 \times 10^6$ ml) compared to the in-domain UKT cohort ($n = 62$, total volume $\approx 8.93 \times 10^6$ ml). Across methods, DSC values frequently clustered around 0, with higher mean FPV@last and mean FNV@last observed at UKT relative to UM. Analysis of lesion topology revealed that 51.6% of lesions at UKT were disappearing, compared with 18.5% at UM, and UKT included merging cases (1.4%) absent at UM (appendix p 13). Additional center-stratified and patient-level analyses are provided in the appendix (pp 14–17).

Figure 5 illustrates representative qualitative results comparing baseline, follow-up, and prediction masks produced by the baseline model. In the exemplary case (top row), the model correctly identifies interval changes in the lung that are visible on the follow-up scan, demonstrating sensitivity to subtle temporal differences. In contrast, the second case shows that the model fails to capture lesion extent present in the mediastinum while simultaneously producing spurious detections in anatomically incorrect regions (red arrows).

Figure 5: Task 2. Qualitative examples compare baseline and follow-up images with their corresponding ground-truth masks (green dots) and the masks predicted by the baseline model (red dots). The top row (A) presents an exemplary case in which the model correctly localizes interval changes in the lung at follow-up. The bottom row (B) illustrates a challenging case of false positive (FP) and false negative (FN) where lesions resolve from baseline to follow-up, highlighting typical failure modes of the model in the mediastinum.

Discussion

We present the results of the autoPET/CT IV challenge: Automated Lesion Segmentation in Whole-Body PET/CT and Longitudinal CT - The human frontier. The challenge focused on two clinically relevant tasks:

(1) interactive lesion segmentation with incremental human guidance, and (2) longitudinal lesion tracking across follow-up scans.

Task 1 results highlight the effectiveness of human-in-the-loop strategies in improving segmentation quality. The top-performing model achieved a mean DSC@last of 0.74, with progressive improvements observed as the number of simulated clicks increased, reflecting reduced FNV. However, FPV increased slightly with more clicks, particularly for the PSMA-UKT subgroup, indicating a sensitivity-specificity trade-off. This may be related to reduced detectability of small or low-contrast lesions, where missing true small lesions during interaction can increase the likelihood of nonspecific or spurious detections, inflating FPV. Although sensitivity for small and low-contrast lesions was generally lower across all algorithms, uptake-based sensitivity remained consistently high (> 0.6) for all subgroups except PSMA-UKT, indicating robust detection in these cases. Nevertheless, mis-segmentation of small or low-uptake lesions highlights the continuing difficulty of distinguishing pathological uptake from physiological background. Performance varied across tracers and centers, consistent with previous challenge editions^{8,9}. UKT data showed higher DSC and lower FNV compared with LMU, whereas LMU scans demonstrated lower FPV, suggesting that domain-specific differences such as scanner protocols and tracer distribution continue to influence algorithm generalizability.

Qualitative analysis revealed systematic error patterns in regions with low contrast, high physiological uptake (abdominal organs), and clusters of small or ambiguous lesions. Interactive clicks reduced some FN errors, but others persisted across interaction levels, suggesting intrinsic limitations of current architectures in handling subtle lesion presentation.

Methodologically, the ensemble model combining all submissions and baselines achieved higher robustness than individual methods, indicating complementary algorithmic strengths. The nnU-Net baseline remained competitive,

highlighting the strong performance of established frameworks. Compared with the top-performing fully automated method from autoPET III, interactive approaches consistently achieved higher performance, suggesting that human-in-the-loop refinement can mitigate key failure modes in challenging PET imaging scenarios. In terms of DSC, gains were larger for PSMA tracers (+ 18.8%) than FDG (+ 6.5%), likely reflecting improved correction in higher-contrast lesions. Despite incremental improvements with human interaction and ensemble integration, no single approach fully resolved challenges related to small lesion detection, high physiological uptake, or domain variability.

From a clinical perspective, results suggest that interactive segmentation can enhance the accuracy and efficiency of lesion delineation in PET/CT imaging. Across tracers and subgroups, interactive methods consistently outperformed fully automated solutions, particularly in challenging cases such as PSMA-UKT with small lesion burden. Interactive refinement could also accelerate dataset annotation by allowing experts to focus on verifying complex cases. Ultimately, these findings highlight the potential of interactive segmentation to support clinical workflows by reducing expert workload while maintaining high lesion detection sensitivity.

Several limitations should be considered. First, simulated clicks represent a simplified interaction paradigm and do not capture the richness of expert input, which may involve scribbles, coarse markings, or context-dependent strategies. Prospective evaluation with clinician-provided interactions is therefore needed to assess effects on diagnostic confidence, annotation efficiency, and inter-observer variability. Second, the static interaction model does not capture dynamic human-algorithm co-adaptation, where sequential feedback could guide model uncertainty and reduce false positives. Third, persistent errors in small or low-uptake lesions suggest limitations of imaging-only approaches; future work could explore multimodal frameworks integrating clinical context or structured reports. Finally, standardizing interaction protocols (e.g., predefined SUV thresholds) across centers may improve generalizability in heterogeneous clinical environments.

Task 2 emphasized the difficulty of segmenting lesion evolution over time. The best-performing algorithm achieved a mean DSC@last of 0.38 at the lesion level, substantially lower than Task 1, reflecting the complexity of accurately segmenting evolving lesions (i.e. changing size, appearance, and location). Patient-level evaluation improved absolute DSC values, suggesting that aggregating performance mitigates individual lesion errors. Temporal priors, including baseline lesion locations (pre-simulated clicks), were beneficial but insufficient to fully capture morphological changes, lesion regression, or the emergence of new lesions. Cross-center analysis showed higher performance on the out-of-domain UM cohort than the in-domain UKT cohort, likely reflecting differences in lesion burden and the higher frequency of disappearing lesions in UKT, which contributed to increased false positives. Especially, the best-performing solution demonstrated stable performance across centers. This approach explicitly modeled temporal context and leveraged longitudinal pretraining, which may facilitate more robust lesion tracking across heterogeneous settings. Overall, results indicate that lesion complexity, beyond dataset imbalance, strongly influences longitudinal segmentation performance.

Qualitative analysis supported these findings: the model reliably captured stable lesions across timepoints but generated spurious detections when lesions resolved or disappeared, highlighting limitations in modeling temporal lesion dynamics.

Clinically, results demonstrate that longitudinal lesion tracking is feasible and may enable more objective assessment of tumor burden during treatment monitoring and response evaluation, where consistent lesion identification across timepoints is essential.

Several limitations should be noted. Evaluation relied on lesion-level multi-class masks, implicitly framing the problem as lesion-level correspondence across timepoints. While this enables fine-grained tracking, it may overemphasize exact lesion continuity and penalize clinically plausible scenarios such as lesion resolution, fragmentation, or merging, which are common in treatment response. The discrepancy between lesion-level and patient-level performance further suggests that strict lesion-wise metrics can obscure meaningful global trends in tumor burden. Additionally, tumor burden is more appropriately represented by cumulative volumetric measures rather than mean-based summaries (FPV, FNV), which may not reflect clinical workload. Architecturally, current longitudinal designs remain limited in their

ability to explicitly reason about lesion appearance, disappearance, and transformation. Reliance on baseline-driven temporal priors may bias models toward persistence assumptions, leading to spurious detections when lesions regress or vanish. Future work could explore architectures that disentangle lesion correspondence from segmentation, incorporate uncertainty in lesion identity over time, or explicitly model longitudinal events such as emergence and resolution.

Methods

This manuscript adheres to the BIAS guidelines¹⁸ for transparent reporting of biomedical image analysis challenges.

Challenge Design and Goals

The autoPET/CT-IV Challenge was organized as part of MICCAI 2025 and supported by the European Society of Radiology (ESR) and the European Society for Hybrid, Molecular and Translational Imaging (ESHI-MT). The challenge was hosted on the Grand Challenge platform, which executed all submissions in isolated Docker environments to ensure standardized evaluation on hidden test data. Building on three prior editions, autoPET/CT-IV introduced an expanded scope centered on interactive, human-in-the-loop lesion segmentation in PET/CT and longitudinal CT imaging.

The challenge evaluated how effectively algorithms leverage simulated user interactions and perform across multicenter, multi-tracer, and multi-timepoint settings. Task 1 assessed how models incorporate simulated user interactions, represented as incremental foreground (lesion) and background (anything else) clicks (up to ten each). Task 2 required lesion segmentation from a single click placed at the lesion center.

Further details regarding the Challenge organization and infrastructure, Participation policies, and Submitted algorithms are provided in the appendix (pp 20–23).

Datasets

Task 1: Single-staging whole-body PET/CT

The Task 1 training dataset includes 1,014 anonymized whole body 18F-FDG PET/CT scans from 900 patients acquired at the University Hospital Tübingen (UKT)¹⁹ and 597 18F-, 68Ga-PSMA PET/CT studies from 378 patients drawn from Ludwig-Maximilians-University (LMU) Munich. The dataset is publicly available in DICOM^{11,12} and NIfTI^{20,21} formats under CC-BY-NC 4.0 license. Acquisitions were performed using multiple scanner types (Table 1). The multicenter and multitracer dataset introduces heterogeneity in acquisition parameters, reflecting real-world clinical variability. Further acquisition details can be found in¹⁹. The held-out test set includes 200 PET/CT studies, stratified across radiotracers and centers (50 cases each from FDG-LMU, FDG-UKT, PSMA-LMU, and PSMA-UKT). Each case consists of one 3D

PET volume, one corresponding 3D CT volume, one 3D binary mask of manually segmented tumor lesions on the PET image, and pre-generated foreground and background clicks to simulate progressive human guidance. Images were preprocessed by converting PET to SUV and resampling CT to match PET spatial resolution. Lesions were annotated by experienced radiologists following a two-stage protocol involving lesion localization and manual segmentation.

Task 2: Longitudinal whole-body CT

The training cohort comprises paired baseline staging and post-therapy follow-up whole-body CT scans from 300 melanoma patients, acquired at the UKT and publicly available in NIfTI format¹³. Acquisition details are depicted in²². Two expert radiologists independently annotated lesions across timepoints using a standardized protocol consisting of CT-based lesion identification, manual axial delineation, and temporal correspondence matching between baseline and follow-up scans. Each case includes at least a 3D baseline (BL) CT volume, the corresponding follow-up (FU) volume, 3D multi-label lesion masks for both timepoints, and a JSON file containing a single center click per lesion to simulate reader annotation in the FU scan. The test set includes 257 longitudinal CT cases with 62 cases from UKT and 195 from University Medical Center Mainz (UM).

Patient characteristics for both tasks are summarized in Table 1. Race and ethnicity data were not systematically collected or recorded in this study and therefore were not analysed or reported.

Assessment Method

Missing values were excluded from average metric computations to prevent bias from incomplete submissions. Evaluation code and dockerized inference environments were publicly released (<https://github.com/lab-midas/autoPETCTIV>) to ensure fairness and transparent benchmarking.

Task 1: Single-staging whole-body PET/CT

Segmentation quality was assessed at the patient level using the Dice Similarity Coefficient (DSC), False Positive Volume (FPV), and False Negative Volume (FNV), capturing overlap accuracy, incorrectly identified lesion volume in non-lesion uptake regions (bladder, spleen, kidney, liver, etc), and missed lesions, respectively. Because the task incorporated simulated human interactions, metrics were also evaluated across ten sequential correction steps. The resulting area-under-the-curve (AUC) quantified improvement across interaction steps. Metric values were averaged across centers (UKT, LMU) and tracers (PSMA, FDG) to account for acquisition variability. Final rankings combined performance after the last interaction (DSC@last, FPV@last, FNV@last) with interaction-efficiency metrics (AUCDSC, AUC-FPV, AUC-FNV), using the following weights: DSC@last (0.25), FPV@last (0.125), FNV@last (0.125), AUC-DSC (0.25), AUC-FPV (0.125), AUC-FNV (0.125). Ranking robustness was assessed using non-parametric bootstrap resampling ($n = 1000$ iterations)²³.

Exploratory post-challenge analysis investigated ensemble combinations of all submissions and baselines using pixelwise majority voting. We further analyzed FP dynamics in relation to lesion count,

lesion volume, and Standardized Uptake Value (SUV). Moreover, lesion-wise sensitivity, defined as the proportion of ground-truth lesions that were correctly detected by the model, was evaluated as a function of lesion size and maximum standardized uptake value (SUV max), with lesions stratified into interdecile bins by dataset domain to identify systematic failure modes. Detection was defined by spatial overlap between a ground-truth lesion and any predicted lesion component.

Additional qualitative analyses examined shared error patterns across algorithms at different interaction levels, focusing on selected test cases with high uptake values, lesion counts, or lesion volumes.

Task 2: Longitudinal whole-body CT

Performance was measured for the single interaction step using DSC@last, FPV@last, and FNV@last computed at the lesion level. Rankings used a weighted combination of DSC@last (0.50), FPV@last (0.25), and FNV@last (0.25), emphasizing accurate lesion detection while penalizing over- and under-segmentation. Supplementary analyses evaluated performance across centers (UKT, UM) to assess institutional robustness. Moreover, metrics were also computed at the patient level to complement lesion-level evaluation, reflecting both individual lesion characteristics and overall disease burden relevant to clinical decision-making.

No ensemble model was developed in this case, as only two submissions were available, limiting the heterogeneity of predictions. Instead, representative cases with and without FP and FN were analyzed qualitatively.

Declarations

Data availability

The dataset is publicly available under a CC BY-NC 4.0 license in both DICOM^{11,12} and NIfTI^{20,21} formats for Task 1, and in NIfTI¹³ format for Task 2.

Declaration of interests

RAW has received speaker honoraria from Novartis. KJ is one of the founders and CEOs of Olatu Medical Technologies GmbH. The work is not linked to the company's activities. The remaining authors declare no competing interests.

Authors contributions

POMC, JD, ZM, NR, TK, APS, SG, KJ, BS, AM, ATS, MPF, LKSS, TL, JHM, AH, DG, AO, AW, FP, AH, KK, RAW, CCC, MI, TK: organization of the challenge, preparation of training and test data, contribution of software,

data analysis, and drafting of the paper. JH, YH, LX, QT, YK, MR, FI, KHM, ZM, SP, LM, YX, KG, MM, AQ: participating teams, contribution in critical revision of the paper.

Acknowledgments

This project was supported by the University Hospital Tübingen and conducted under Germany's Excellence Strategy EXC-Numbers EXC 2064/1-390727645 and EXC 2180/1-390900677. This work was funded by the Federal Ministry of Research, Technology and Space (BMFTR) under project 01KD2417A, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project 428216905 / SPP 2177, and by the Adolf Leuze foundation. MR is funded through a Google PhD Fellowship. For AH, this publication is part of the project OncoChange with file number 21121 of the research programme Veni - Applied and Engineering Sciences, which is (partly) financed by the Dutch Research Council (NWO).

References

1. Thie JA. Understanding the Standardized Uptake Value, Its Methods, and Implications for Usage. *Journal of Nuclear Medicine*. 2004;45(9):1431–4.
2. Boellaard R, Delgado-Bolton R, Oyen WJ, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging*. 2015;42(2):328–54.
3. Adelsmayr G, Janisch M, Kaufmann-Buhler AK, Holter M, Talakic E, Janek E, et al. CT texture analysis reliability in pulmonary lesions: the influence of 3D vs. 2D lesion segmentation and volume definition by a Hounsfield-unit threshold. *Eur Radiol*. 2023;33(5):3064–71.
4. Foster B, Bagci U, Mansoor A, Xu Z, Mollura DJ. A review on segmentation of positron emission tomography images. *Comput Biol Med*. 2014;50:76–96.
5. Guezennec C, Bourhis D, Orlhac F, Robin P, Corre JB, Delcroix O, et al. Inter-observer and segmentation method variability of textural analysis in pre-therapeutic FDG PET/CT in head and neck cancer. *PLoS One*. 2019;14(3):e0214299.
6. Parkinson C, Foley K, Whybra P, Hills R, Roberts A, Marshall C, et al. Evaluation of prognostic models developed using standardised image features from different PET automated segmentation methods. *EJNMMI Res*. 2018;8(1):29.
7. Pfaehler E, Burggraaff C, Kramer G, Zijlstra J, Hoekstra OS, Jalving M, et al. PET segmentation of bulky tumors: Strategies and workflows to improve inter-observer variability. *PLoS One*. 2020;15(3):e0230901.
8. Gatidis. S, Früh. M, Fabritius. MP, Gu. S, Nikolaou. K, Fougère. CL, et al. Results from the autoPET challenge on fully automated lesion segmentation in oncologic PET/CT imaging. *Nature Machine Intelligence*. 2024;6(11):1396–405

9. Dexl J, Gatidis S, Fruh M, Jeblick K, Mittermeier A, Stuber AT, et al. AutoPET Challenge on Fully Automated Lesion Segmentation in Oncologic PET/CT Imaging, Part 2: Domain Generalization. *J Nucl Med.* 2026;67(3):481–8.
10. Dexl. J, Jeblick. K, Mittermeier. A, Schachtner. B, Stüber. AT, Topalis. J, et al. The autoPET3 Challenge: Automated Lesion Segmentation in Whole-Body PET/CTMultitracer Multicenter Generalization. *arXiv.* 2026.
11. Gatidis S, Kuestner T. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions (FDG-PET-CT-Lesions). *The Cancer Imaging Archive.* 2022;Dataset.
12. Jeblick K, Schachtner B, Mittermeier A, Dexl J, Wesp PaK, T., Gatidis SaF, M., et al. A whole-body PSMA-PET/CT dataset with manually annotated tumor lesions. *The Cancer Imaging Archive.* n.d.;Dataset.
13. Küstner T, Peisen F, Gatidis S, Wagner A, Megne O, Othman A, et al. Longitudinal-CT. 2025.
14. Ronneberger O, Fischer P, Brox T, editors. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*; 2015.
15. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods.* 2021;18(2):203–11.
16. Munk A, Ambsdorf J, Llambias S, Nielsen M. Amaes: Augmented masked autoencoder pretraining on public brain mri data for 3d-native segmentation. *arXiv preprint arXiv:240800640.* 2024.
17. Beck M, Pöppel K, Spanring M, Auer A, Prudnikova O, Kopp M, et al. xlstm: Extended long short-term memory. *Advances in Neural Information Processing Systems.* 2024;37:107547–603.
18. Maier-Hein L, Reinke A, Kozubek M, Martel AL, Arbel T, Eisenmann M, et al. BIAS: Transparent reporting of biomedical image analysis challenges. *Med Image Anal.* 2020;66:101796.
19. Gatidis S, Hepp T, Fruh M, La Fougere C, Nikolaou K, Pfannenberger C, et al. A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions. *Sci Data.* 2022;9(1):601.
20. Gatidis S, Küstner T, Ingrisich M, Hepp T, Früh M, Nikolaou K, et al. FDG-PET-CT-Lesions. 2022;Dataset.
21. Gatidis S, Küstner T, Ingrisich M, Hepp T, Früh M, Nikolaou K, et al. PSMA-FDG-PET-CT-Lesions. 2024;Dataset.
22. Gatidis S, Peisen F, Wagner A, Choudja POM, Othman A, Sanner A, et al. A longitudinal whole-body CT dataset with manually annotated tumor lesions. *Sci Data.* 2026;13(1).
23. Wiesenfarth M, Reinke A, Landman BA, Eisenmann M, Saiz LA, Cardoso MJ, et al. Methods and open-source toolkit for analyzing and visualizing challenge results. *Sci Rep.* 2021;11(1):2369.

Table

Table 1 is available in the Supplementary Files section.

Figures

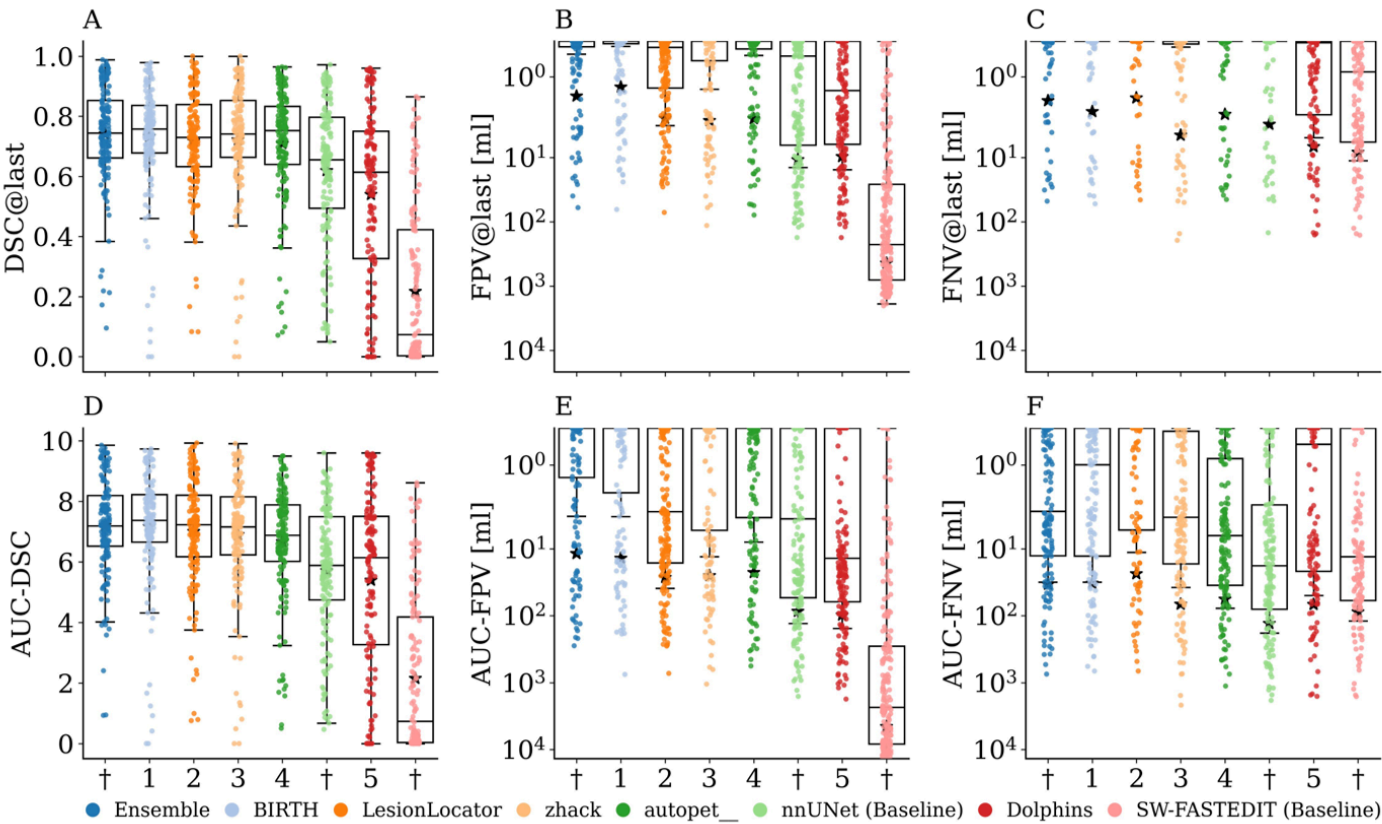


Figure 1

Task 1. Performance distribution of all ranked algorithms across the six evaluation metrics: (A) Dice Similarity Coefficient (DSC), (B) False Positive Volume (FPV), and (C) False Negative Volume (FNV) at the last interactive step (i.e. 10 foreground and 10 background clicks), and (D-F) the respective area-under-the-curve (AUC). Bars are ordered from left to right according to the official challenge ranking. The ensemble, nnU-Net, and SW-FASTEDIT baselines are out of competition (†). The star symbolizes the average performance.

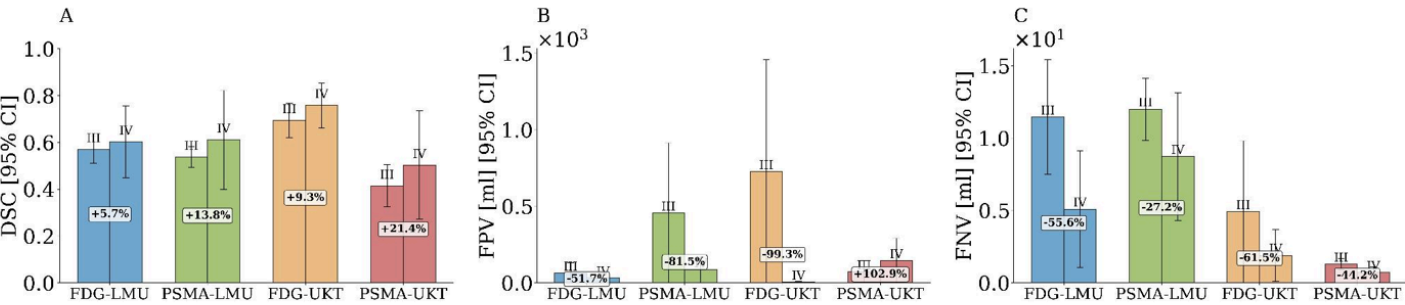


Figure 2

Task 1. Comparison of all submissions from autoPET/CT-IV (IV) to the models from autoPET III (III). Percentage improvement (from III to IV) is displayed between paired bars. Overall, autoPET IV algorithms demonstrate superior segmentation quality with reduced FPV (except in PSMA-UKT) and FNV across all cohorts, indicating enhanced generalizability. Error bars represent the 95% confidence interval (CI).

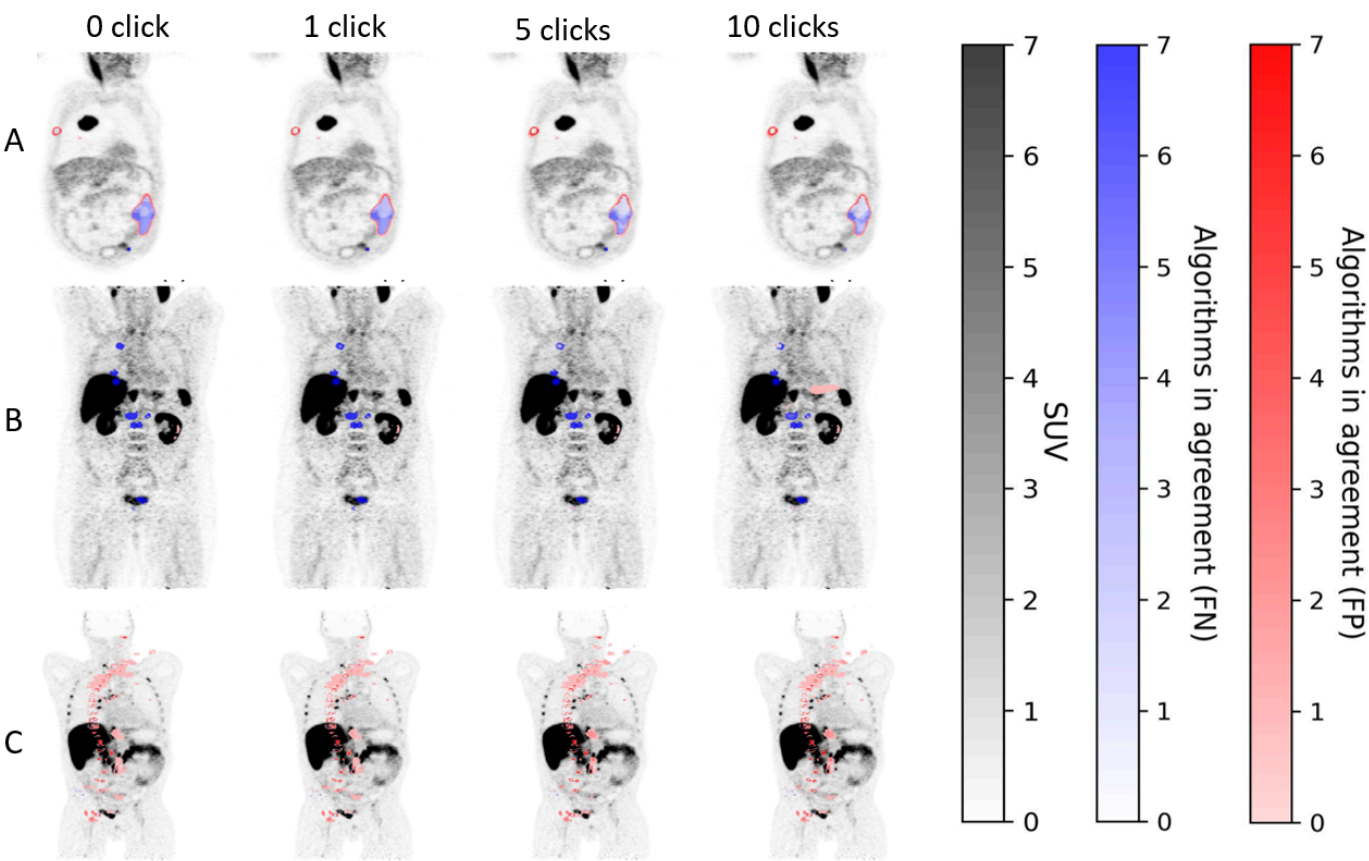


Figure 3

Task 1. Qualitative analysis of common segmentation error maps at different levels of user interaction. Segmentation errors from each algorithm are overlaid on PET images for increasing numbers of foreground and background clicks (0, 1, 5, and 10). False positive errors (FP) are shown in red and false negative (FN) errors in blue. Colour intensity reflects the number of individual algorithms (excluding the Ensemble model) that misclassify a given voxel, ranging from 0 (all algorithms correct at that voxel) to 7 (seven algorithms incorrect).

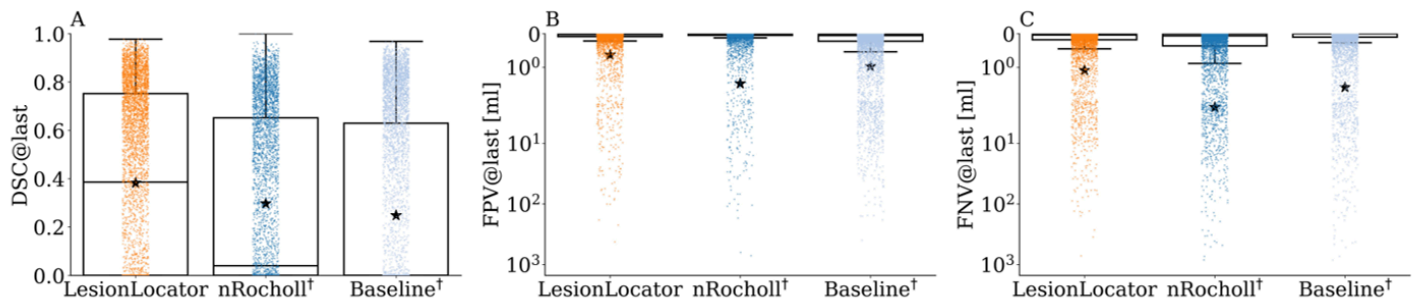


Figure 4

Task 2. Lesion-level segmentation performance at the follow-up (FU) time point for all ranked algorithms across the three evaluation metrics (A, B, C). Bars are ordered from left to right according to the official challenge ranking. The nRocholl team model and LongiSeg baseline are out of competition (†). The star symbolizes the average performance.

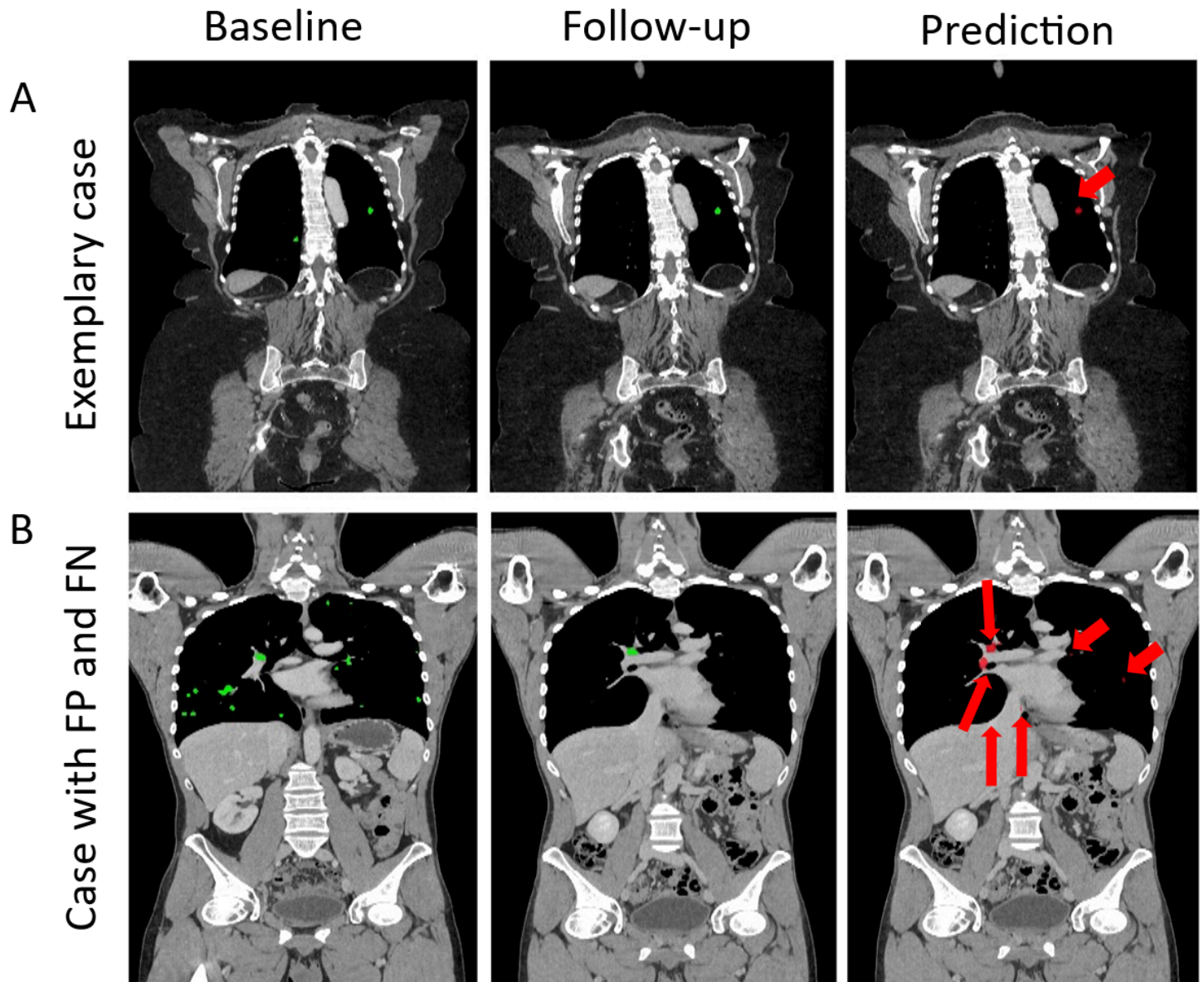


Figure 5

Task 2. Qualitative examples compare baseline and follow-up images with their corresponding ground-truth masks (green dots) and the masks predicted by the baseline model (red dots). The top row (A) presents an exemplary case in which the model correctly localizes interval changes in the lung at follow-up. The bottom row (B) illustrates a challenging case of false positive (FP) and false negative (FN) where lesions resolve from baseline to follow-up, highlighting typical failure modes of the model in the mediastinum.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementmaterials.pdf](#)
- [Table1.docx](#)