The autoPET challenge: Towards fully automated lesion segmentation in oncologic PET/CT imaging

Sergios Gatidis (Sergios.Gatidis@med.uni-tuebingen.de)
University Hospital Tuebingen
Marcel Früh
University Hospital Tübingen
Matthias Fabritius
LMU University Hospital
Sijing Gu
LMU University Hospital
Konstantin Nikolaou
University Hospital Tübingen
Christian La Fougère
University Hospital Tübingen
Jin Ye
Shanghai AI Lab
Junjun He
Shanghai AI Lab
Yige Peng
University of Sydney https://orcid.org/0000-0001-5549-2688
Lei Bi
School of Computer Science
Jun Ma
University of Toronto
Bo Wang
Peter Munk Cardiac Centre
Jia Zhang
United Imaging Healthcare
Yukun Huang
United Imaging Healthcare
Lars Heiliger
University Hospital Essen
Zdravko Marinov
Karlsruhe Institute of Technology https://orcid.org/0000-0003-0373-3958
Rainer Stiefelhagen

Karlsruhe Institute of Technology
Jan Egger
University Hospital Essen
Jens Kleesiek
Institute for AI in Medicine, University Medicine Essen https://orcid.org/0000-0001-8686-0682
Ludovic Sibille
Subtle Medical
Lei Xiang
Subtle Medical
Simone Bendazolli
KTH Royal Institute of Technology
Mehdi Astaraki
KTH Royal Institute of Technology
Bernhard Schölkopf
Max Planck Institute for Intelligent Systmes
Michael Ingrisch
University Hospital, LMU Munich https://orcid.org/0000-0003-0268-9078
Clemens Cyran
University Hospital, LMU Munich
Thomas Küstner
University Hospital Tübingen

Article

Keywords: Machine Learning, Challenge, PET/CT, FDG, oncology, segmentation

Posted Date: June 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2572595/v1

Abstract

We describe the results of the autoPET challenge, a biomedical image analysis challenge aimed to motivate and focus research in the field of automated whole-body PET/CT image analysis. The challenge task was the automated segmentation of metabolically active tumor lesions on whole-body FDG-PET/CT. Challenge participants had access to one of the largest publicly available annotated PET/CT data sets for algorithm training. Over 350 teams from all continents registered for the autoPET challenge; the seven best-performing contributions were awarded at the MICCAI annual meeting 2022. Based on the challenge results we conclude that automated tumor lesion segmentation in PET/CT is feasible with high accuracy using state-of-the-art deep learning methods. We observed that algorithm performance in this task may primarily rely on the quality and quantity of input data and less on technical details of the underlying deep learning architecture. Future iterations of the autoPET challenge will focus on clinical translation.

Introduction

Recent advances in computational medical image analysis – in particular the introduction of deep learning methods – have led to substantial progress in numerous medical image analysis tasks including segmentation, regression and classification tasks. As part of this rapid development, medical image analysis challenges have played a crucial role by identifying relevant tasks, motivating and coordinating engagement, defining benchmarks and – perhaps most importantly – providing publicly available labeled data for algorithm development.

Several prominent examples of medical image analysis challenges, such as The Medical Segmentation Decathlon[1], the BRATS challenge[2] or the RSNA Pediatric Bone Age Challenge[3] illustrate the immense impact of such initiatives on their respective fields of research and application.

Most medical imaging challenges focus on the analysis of normal anatomy or the analysis of pathologies in defined anatomic regions, limiting the scope and complexity as well as the amount of required training data. In comparison, computational analysis of whole-body oncologic examinations, as acquired by PET/CT, is associated with higher complexity due to the multimodal nature of the underlying data, the large anatomical coverage, and the high morphological variability of oncologic pathologies. Furthermore, the generation of training labels on oncologic whole-body examinations requires a high level of clinical expertise and can only be performed by experienced medical imaging specialists. These factors contribute to delayed progress in the field of computational whole-body oncologic image processing, specifically regarding whole-body PET/CT imaging. Few studies have reported the development and application of automated PET/CT analysis – specifically automated tumor lesion segmentation – in the past. In these studies, a variety of methodological approaches have been proposed ranging from simple, threshold-based segmentation algorithms[4] to state-of-the-art deep learning methods[5] or combinations thereof[6]. While these studies clearly demonstrate the technical feasibility of automated PET/CT image analysis, the comparison and reproducibility of methods is limited due to the use of proprietary data and algorithms. A recent medical image analysis challenge (HECKTOR

challenge)[7] on automated PET/CT lesion segmentation in the head/neck region demonstrated in an anatomically restricted scenario, how combined efforts by the research community can advance this field in a specified direction.

Automation of the image analysis process in oncologic whole-body PET/CT data is of high interest. Quantitative analysis of PET/CT data requires segmentation of tumor lesions which is time-consuming and labor-intensive, thus associated with high effort and cost. This prevents wide-spread clinical adoption of quantitative image analysis beyond study settings. Automation of this process can thus potentially allow for integration of quantitative PET/CT analysis in routine clinical workflow supporting diagnostic and therapeutic decisions.

To advance the field of automated oncologic PET/CT analysis and to address the existing shortcomings in this area, we conducted the autoPET challenge. The primary challenge task was the automated segmentation of metabolically active tumor lesions in whole-body FDG-PET/CT. To this end, a multi-center database of 1164 oncologic whole-body PET/CT datasets (1014 public training samples and 150 private test samples) with manually segmented tumor lesions was composed. The training dataset is publicly available at The Cancer Imaging Archive (TCIA)[8] and has been previously described in detail[9].

In this work we present the content and results of the autoPET challenge that was conducted as part of MICCAI 2022 aiming to (1) motivate and focus research in the field of automated PET/CT image analysis (2) provide a platform for algorithm comparison and reproduction and (3) document the current state-of-the-art in this field. In addition, we provide analysis on the importance of composition and size of available training data for successful algorithm development.

Results

In the following, we describe the challenge preparation, organization and evaluation following the guidelines for biomedical image analysis challenge reporting (BIAS guidelines)[10]. The public challenge training data set was drawn from the University Hospital Tübingen (UKT). The private challenge test set was partly drawn from the same source (UKT) and partly from the University Hospital of the LMU Munich (LMU).

Challenge Participation

A total number of 359 teams registered for the autoPET challenge including teams from all continents (Fig. 1) with clear geographic concentrations on Asia (61%, mainly China: 41%), Europe (20%) and North America (16%, mainly USA: 13%). As far as disclosed, most participants were affiliated to academic institutions (75%), followed by a smaller group of company employees (12%).

37 teams submitted at least one algorithm to the preliminary challenge phase amounting to 253 total submissions in this phase. In the final challenge phase 18 teams contributed a total of 67 algorithms. The best-performing submission by each team was considered for the challenge leaderboard. The seven

best performing teams were identified as challenge winners – their contributions are described in greater detail as part of this work.

As expected, all final contributions were based on deep learning models. The majority of submitted algorithms relied on a 3D U-Net backbone in combination with a Dice loss. A minority of participants deployed transformer-based architectures or combinations of different architectures (2D and 3D) or used more uncommon loss functions[11] (e.g., focal loss, TopK Dice loss, Lovasz loss, Tversky loss), mostly in combination with a conventional Dice loss. An overview of the technical details is depicted in Fig. 2.

Best performing algorithms

In the following, we provide brief descriptions of the seven best-performing contributions in the order of the final leaderboard followed by individual performance reports. The code for all contributions is publicly available – details are available in the technical papers published by the participating teams and cited below. Overall, the use of a U-Net backbone was a common feature of the best contributions. The additional implementation of rule-based post-processing of algorithm outputs (e. g. threshold-based removal of small connected components from the output segmentation mask) distinguished the top four contributions from the rest of the field. All top-performing teams used both, the PET and CT image volumes as algorithm inputs.

Team Blackbean

The best-performing team chose a deliberately simple approach by using a vanilla U-Net backbone and focusing on ablation studies to identify the best combination of input shape (crop size) and step size during sliding window inference. In addition, a post-processing step was used to minimize the false-positive volume by removing small connected components from the initial algorithm output[12].

Team BDAV

Team BDAV used a combination of self-supervised pre-training (via contrastive learning) and a multi-stage U-Net architecture. The multi-stage U-Net architecture utilized a global segmentation module to conduct coarse tumor segmentation, which was then fed into a local refinement module to reduce the false positives. The multi-stage U-Net model was ensembled with a standard nnUNet model to generate the final prediction[13].

Team FightTumor

This contribution was based on a slightly modified nnUNet model using DiceTopK loss and enhanced data augmentation. In addition, post-processing of the model output was performed by removing small connected components (< 10 voxels) and segmentations in areas with low CT Hounsfield Units (< -1,000 HU)[14].

Team UIH-FL

Team UIH-FL trained a combined 2D and 3D nnUNet model. In addition, they performed post-processing of the model output by removing small connected components (< 4 voxels) and all connected components on the three bottom slices of the predicted pet mask[15].

Team Heiligerl

This contribution was based on an ensemble of an nnUNet-based model and a Swin UNETR. In addition, a classification model was trained to identify negative PET/CT scans without metabolically active lesions, based on maximum intensity projections (MIP), inspired by reading procedures of physicians [16].

Team SM

Team SM proposed a cascaded architecture consisting of a stacked ensemble of low-resolution U-Net models and a subsequent refiner U-Net for high-resolution predictions[17].

Team Flemings

Team Flemings proposed a cascaded architecture consisting of an initial inpainting model to detect and generate lesion-free images, followed by a U-Net-based segmentation, with the residual inpainting image as additional input[18].

nn-Unet (baseline model, out of competition)

To provide a baseline model, the widely used and standardized nn-UNet framework[19] was used with the default settings using PET and CT volumes as input. The trained baseline model is publicly available under https://github.com/lab-midas/autoPET.

Ensemble model (out of competition)

Based on the predictions of these above-described best performing algorithms, including the baseline model, an ensemble model output was computed by pixel-wise majority voting.

Overall, the performance metrics of the best performing teams were slightly different with respect to all three metrics (Fig. 3, A): Dice score (capturing consistency between foreground predictions and manual masks), false negative volume (capturing total volume of missed lesions), and false positive volume (capturing false-positive segmentations of physiologic tracer uptake). The mean Dice score ranged between 0.74 and 0.79, the mean false negative volume between 0.5 and 1.5 ml and the false positive volume between 2.1 and 9.5 ml. When assessing algorithm performance separately for the two data sources (UKT and LMU) of the multicentric training set, we observed that mean Dice scores were overall markedly higher for UKT test data (ranging between 0.8 and 0.88) compared to LMU test data (ranging between 0.6 and 0.7). Mean false negative volumes and false positive volumes were slightly higher for LMU data compared to UKT data (false negative volumes UKT: 0.3 to 1.7 ml, false negative volumes LMU: 0.9 to 2.3 ml; false positive volumes UKT: 1.5 to 5.4 ml, false positive volumes LMU: 3.2 to 20.3 ml).

The best performing team (Blackbean) ranked first regarding the mean dice score and the mean false negative volume and second regarding the mean false positive volume. Interestingly, the provided baseline nnUNet model showed a good overall performance ranking – out of competition – second with respect to the mean Dice score and seventh with respect to the mean false positive and false negative values (Fig. 3, A).

The ensemble prediction (out of competition) based on the top performing contributions and the baseline model showed a superior performance compared to all participating teams with the highest overall mean Dice score (0.81), the second lowest mean false negative volume (0.71 ml) and the lowest mean false positive volume (1.6 ml) (Fig. 3, A).

Typical qualitative examples of model performance and error cases are given in Fig. 4. In general, false positive segmentations mainly occurred in areas of atypical physiological tracer uptake (e. g. unusually large urinary bladder, brown adipose tissue) while tumor lesions adjacent to physiological tracer uptake were more often missed.

Impact of training data composition on algorithm performance

To better understand external factors influencing algorithm performance in general we performed additional ablation studies using the baseline model with different sizes and compositions of training data.

As could be expected, we observed an overall increase in segmentation performance with increasing numbers of training data reflected by increasing Dice scores and decreasing false positive and false negative volumes (Fig. 3, B). Interestingly, in contrast to this overall tendency, Dice scores on LMU test data showed no increase and even slightly decreased with higher numbers of training data, probably as a sign of overfitting to the UKT training data distribution.

In addition, we assessed the impact of the input data composition on algorithm performance. In addition to PET and CT volumes that were also used within the challenge, we added CT-based anatomical organ labels as a potential third input.

Regarding the composition of training data, we observed the highest segmentation performance in terms of Dice scores when using all three inputs (PET, CT and anatomical labels) on both, UKT and LMU data (Fig. 3, B). On UKT data, using all three inputs also gave lower false positive and false negative volumes. On LMU data, the results regarding composition of data and false positive/negative volumes were inconclusive; however, using only PET data resulted in markedly higher false positive volumes on LMU test data.

Figure 5 provides qualitative examples of test data sets and associated segmentation results for test data drawn from UKT and LMU. In agreement with the quantitative results, we qualitatively observed a

larger mismatch between manual and automated tumor lesion segmentation on LMU data (Fig. 5). In general, tumor volumes were locally overestimated on LMU test data explaining the lower Dice scores on LMU test data compared to UKT test data. Also agreeing with the quantitative results, with respect to false positive and false negative volumes, we did not observe any obvious qualitative differences between LMU and UKT test data.

Discussion

In this work we introduce the autoPET challenge on automated PET/CT lesion segmentation – organized as a MICCAI challenge in 2022 – and present its results as well as the results of further analyses to pave the way for a clinical adoption of automated PET/CT image analysis.

The main technical scope of this challenge was the automated segmentation of metabolically active tumor lesions in whole-body FDG PET/CT scans. The best performing contributions demonstrated that this basic and important task can be performed using state-of-the-art deep learning methodology with high overall accuracy.

Interestingly, algorithm performance did not depend in a relevant way on technical details of the deep learning architecture, and the provided nnUNet-based baseline model already performed among the top contributions. Furthermore, an ensemble model of the top-performing algorithms showed the best overall performance. These observations are in line with more general results from machine learning challenges indicating that ensembling of many different algorithms can be superior to optimization of a single algorithm [20].

In contrast, the size and composition of training data had a substantial effect on algorithm performance. First, we observed a slight but relevant increase in lesion segmentation accuracy when using PET and CT data as input compared to a PET-only input indicating that anatomical and morphological information is useful for lesion segmentation. Overall, segmentation accuracy also increased with increasing the size of the training set. However, this effect was not uniform between test data from UKT (same as training distribution) and test data from LMU (different from training distribution): On UKT test data, the increase in training data resulted in marked increase in Dice scores and decrease in false negative volumes with a plateau at around 800-1,000 training samples. On LMU test data however, while false negative volumes also increased with increasing training examples, no improvement and even a slight decline in Dice scores was observed. False positive volumes were relatively low in both test data sets independent of the size of the training set. These results indicate that the generalizability of algorithms trained in a single institution is limited and that a reduction in segmentation performance can be expected when applied to data from different sources, e.g., different scanners of hospitals. This drop in performance is not catastrophic but rather related to different localization of the tumor margins – false positive and false negative volumes were interestingly not worse on the external test data. These results motivate us to place our focus on the topics of robustness and generalization for the next iteration of the autoPET challenge.

The autoPET 2022 is the first, important step towards a long-term goal of fully automated, quantitative oncologic PET/CT image analysis. A number of tasks – beyond lesion segmentation in FDG-PET – need to be addressed. For the near future we identify mainly two: (i) generalization of PET lesion segmentation to different tracers and to different environments (tumor types, scanners, hospitals etc.) and (ii) segmentation of lesions that are only visible on CT due to low or missing tracer uptake. We aim to include these tasks as part of the next iterations of the autoPET challenge.

Materials and Methods

Challenge Mission and Task

The mission of the autoPET challenge is to motivate and focus research in the field of automated PET/CT image analysis, to provide a platform for algorithm comparison and reproduction and to document the current state-of-the-art in this field.

The autoPET challenge task – fully automated segmentation of metabolically active tumor lesions – is a crucial first step towards objective and quantitative oncologic diagnosis, staging and therapy response assessment in whole-body FDG-PET/CT. This task can be performed manually in principle but – depending on tumor spread – can be associated with enormous effort by experts. As a result, lesion segmentation is not performed routinely in clinical settings. Automation of this task will strongly support clinical implementation of PET/CT lesion segmentation and quantitative analysis.

We consider the autoPET challenge 2022 to be the first part of a series of challenges aiming to gradually address increasingly complex aspects of automated PET/CT analysis including detection and segmentation of tumor lesions on CT data, extension to other PET tracers, lesion phenotyping and the analysis of longitudinal imaging studies.

Challenge Organization and Infrastructure

The autoPET Challenge was conducted in 2022 as a MICCAI (Medical Imaging Computing and Computed Assisted Intervention Society) - registered challenge and in cooperation with the European Society of Hybrid, Molecular end Translational Imaging (ESHI-MT). The organizing team consisted of radiologists and medical data scientists from the University Hospital Tübingen (UKT) in Tübingen, Germany and the Ludwig-Maximilian-University (LMU) Hospital in Munich, Germany.

The challenge proposal was submitted to MICCAI in December 2021 and – after undergoing a peer-review process – was approved in February 2022. The final challenge proposal is publicly available [21]. The challenge and its results were presented in a Satellite Event at the 25th International Conference on Medical Imaging Computing and Computed Assisted Intervention in September 2022.

The challenge was opened on April 1st, 2022, with the release of all related information and the public training set. During a first submission phase, starting May 3rd, 2022, participants were able to submit

their algorithms to perform technical sanity checks on a small, preliminary private test data set consisting of 5 representative test samples. The second and final submission phase was launched August 1st, 2022, with the activation of the final, private test data set consisting of 150 test samples (Fig. 1).

The technical realization of autoPET 2022 was conducted on the dedicated Grand Challenge platform (grand-challenge.com, Diagnostic Image Analysis Group, Radboud University Medical Center, The Netherlands) as a type-II challenge (i.e., submission of algorithms by participants to run on a private test set) under the URL https://autopet.grand-challenge.org/. Due to the private nature of the test data set, algorithms were submitted to and deployed on the Grand Challenge computing platform via Docker containers. A time limit of 20 minutes per test sample was set for algorithm execution on the available computation resources (1 TPU with 16 GB GPU memory (NVIDIA, Santa Clara, CA, USA), 8 CPUs and 30 GB of CPU memory).

Technical support was provided to challenge participants in the form of a baseline algorithm example and the associated code base as well as detailed description of the submission process on a public online code repository (https://github.com/lab-midas/autoPET) and the challenge website (https://autopet.grand-challenge.org/).

Participation policies

The use of data for algorithm development was restricted to the provided public training data set. No additional data or machine learning models pre-trained on external data were permitted. Members of the organizer's institutes were allowed to participate in the autoPET challenge but were not eligible for awards. The seven best performing contributions according to the challenge leader board (ranking criteria are described below) were announced publicly awarded with monetary prizes (in €: 6,000 for first place, 3,000 for second place, 2,000 for third place and 1,000 for places four to seven). To be eligible for awards, participating teams were required to publish their code and a technical manuscript describing their methodology and results under an open-source license. Two members of each team were invited to contribute as authors to this manuscript.

Datasets

The public training data consisted of 1,014 anonymized oncologic whole-body FDG-PET/CT data sets together with manually generated segmentation labels of metabolically active tumor lesions drawn from the University Hospital in Tübingen (UKT) (Fig. 6, A). All training data were obtained from a single institution and clinical PET/CT scanner (Biograph mCT, Siemens Healthcare) using a standardized imaging protocol (CT protocol: reference tube current, 200 mAs with automated tube current modulation; tube voltage, 120 kV; i.v. contrast agent injection, 90–120 ml Ultravist 370 (Bayer AG) in the portal-venous phase; slice thickness, 2–3 mm; PET protocol: tracer uptake time, 60 min; injected tracer activity, 300–350 MBq; iterative reconstruction (two iterations, 21 subsets) with Gaussian smoothing (2 mm full width at half-maximum); reconstructed voxel size, 2 x 2 x 3 mm^3). The training data set is publicly available at The Cancer Imaging Archive (TCIA)[8] and has been previously described in detail[22].

The private test data set consisted of 150 anonymized oncologic whole-body FDG-PET/CT data sets together with manually generated segmentation labels of metabolically active tumor lesions (Fig. 6, A). 100 of the 150 training samples were obtained from the same institution (UKT) and acquired with the same imaging protocol as the training data set. The remaining 50 of the 150 training samples were obtained from a different institution (LMU) with variable PET/CT imaging protocols using clinical PET/CT scanners by two vendors (64 TruePoint or Biograph mCT Flow (Siemens Healthineers) or a GE Discovery 690 (GE Healthcare)). These 50 test data sets were acquired using similar protocols (CT protocol: tube current, 100–190 mAs; tube voltage, 120 kV; i.v. contrast agent injection, weight-adapted dose of Ultravist 300 (Bayer AG) or Imeron 350 (Bracco Imaging) in the portal venous phase; slice thickness, 3 mm; PET protocol: tracer uptake time, 60 min; tracer activity, 300–350 MBq; iterative reconstruction (three iterations, 21 subsets) with Gaussian smoothing (2–4 mm full width at half-maximum) reconstructed voxel size, 2.7 x 2.7 x 2–4 x 4 x 5 mm^3).

Only members of the organizing committee had access to the private test data set and its labels.

Both, the training and test data sets included examinations of patients diagnosed with lymphoma, lung cancer or melanoma as well as negative studies (without detectable metabolically active tumor lesion). Training and test populations had a comparable age distribution. 444 of 1,014 training scans (43.7%) and 58 of 150 test scans (38.7%) were of female patients. Regarding the distribution of tumor load, measured by the Metabolic Tumor Volume (MTV) and metabolic tumor activity, measured by the mean Standardized Uptake Value (meanSUV) training and test data showed a good overall agreement (Supplemental Figure S1).

A representative example of a complete data set is provided in the supplemental material (Supplemental Figure S2).

Evaluation and further analyses

Quantitative algorithm performance regarding the challenge task was assessed using three metrics representing different aspects as previously described in [22] (Fig. 6, B). The foreground Dice score was used as an overall metric of agreement between ground truth segmentations and algorithm predictions. In addition, the metrics “false positive volume” and “false negative volume” were used to quantify the erroneous segmentation of healthy tissue and the miss of entire tumor lesions respectively. The false positive volume is defined as the sum of all positive connected components in the prediction that do not overlap with true tumor lesions in the ground truth (i.e., false positive segmentations that are not related to actual tumor lesions). The false negative volume is defined as true positive connected components that do not overlap with positive areas of the prediction (i.e., tumor lesions that were entirely missed).

Challenge submissions were ranked separately for each to these three metrics using their respective means. The final overall rank was derived using the mean of these single rankings (with Dice score being weighted twice) using the Dice score as a tie break. The code used for computation of the challenge metrics is publicly available under https://github.com/lab-midas/autoPET.

To analyze the generalization properties of submitted algorithms, all metrics were also computed separately for the two data sources (UKT and LMU).

To assess the impact of available training data the following additional ablation studies were performed using baseline models based on the standard configuration of the nn-UNet framework[19]: The impact of the number of available training data was assessed by training different versions of the baseline model with 50, 100, 200, 400, 800 or 1,014 randomly drawn training datasets. To assess the impact of additional anatomical information, these baseline models were trained either using only the PET image volumes as model input, or the PET volumes and the corresponding CT volumes or the PET and CT volumes together with CT-based anatomical organ segmentation masks. These organ segmentation masks were derived on using a publicly available pre-trained CT organ segmentation model [23]. This model provides segmentation of 36 anatomical structures including all major organs as well as adipose and lean tissue compartments (Supplemental Fig. 2). The underlying hypothesis for these experiments was that the addition of implicit or explicit anatomical information as input might support the learning process and potentially improve segmentation performance or reduce the number of required training data. It should be noted that the use of anatomical labels was not permitted within the challenge and was only used as part of these additional analyses.

Declarations

Acknowledgements

This project was partly supported by the Leuze Foundation, Owen/Teck, Germany. This project was conducted under Germany's Excellence Strategy–EXC-Number 2064/1–390727645 and EXC 2180/1-390900677. This study is part of the doctoral thesis of Alexandra Kubičkova.

Competing Interests

The Authors declare no Competing Financial or Non-Financial Interests.

Author Contributions

Sergios Gatidis, Marcel Früh, Sijing Gu, Matthias Fabritius, Michael Ingrisch, Clemens Cyran, Thomas Küstner:

Organization of the challenge, Preparation of training and test data, Contribution of software, Data analysis, Drafting of the manuscript

Konstantin Nikolaou, Christian la Fougère, Bernhard Schölkopf:

Scientific and clinical consultation during challenge preparation and data analysis. Critical revision of the manuscript

Jin Ye, Junjun He, Yige Peng, Lei Bi, Jun Ma, Bo Wang, Jia Zhang, Yukun Huang, Lars Heiliger, Zdravko Marinov, Jens Kleesiek, Rainer Stiefelhagen, Jan Egger, Ludovic Sibille, Lei Xiang, Simone Bendazzoli, Mehdi Astaraki:

Members of the best performing participating teams. Contribution of software. Participation in drafting and critical revision of the manuscript.

References

1. Antonelli, M. *et al.* The Medical Segmentation Decathlon. *Nat Commun* **13**, 4128 (2022). https://doi.org:10.1038/s41467-022-30695-9
2. Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging* **34**, 1993-2024 (2015). https://doi.org:10.1109/tmi.2014.2377694
3. Halabi, S. S. *et al.* The RSNA Pediatric Bone Age Machine Learning Challenge. *Radiology* **290**, 498-503 (2019). https://doi.org:10.1148/radiol.2018180736
4. Weisman, A. J. *et al.* Comparison of 11 automated PET segmentation methods in lymphoma. *Phys Med Biol* **65**, 235019 (2020). https://doi.org:10.1088/1361-6560/abb6bd
5. Groendahl, A. R. *et al.* A comparison of fully automatic segmentation of tumors and involved nodes in PET/CT of head and neck cancers. *Phys Med Biol* (2021). https://doi.org:10.1088/1361-6560/abe553
6. Capobianco, N. *et al.* Deep-Learning (18)F-FDG Uptake Classification Enables Total Metabolic Tumor Volume Estimation in Diffuse Large B-Cell Lymphoma. *J Nucl Med* **62**, 30-36 (2021). https://doi.org:10.2967/jnumed.120.242412
7. Oreiller, V. *et al.* Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. *Medical Image Analysis* **77**, 102336 (2022). https://doi.org:https://doi.org/10.1016/j.media.2021.102336
8. Gatidis, S. & Kuestner, T. (The Cancer Imaging Archive (TCIA), 2022).
9. Gatidis, S. *et al.* A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions. *Sci Data* **9**, 601 (2022). https://doi.org:10.1038/s41597-022-01718-3
10. Maier-Hein, L. *et al.* BIAS: Transparent reporting of biomedical image analysis challenges. *Medical Image Analysis* **66**, 101796 (2020). https://doi.org:https://doi.org/10.1016/j.media.2020.101796
11. Ma, J. *et al.* Loss odyssey in medical image segmentation. *Medical Image Analysis* **71**, 102035 (2021). https://doi.org:https://doi.org/10.1016/j.media.2021.102035
12. Ye, J. *et al.* Exploring Vanilla U-Net for Lesion Segmentation from Whole-body FDG-PET/CT Scans. arXiv:2210.07490 (2022). <https://ui.adsabs.harvard.edu/abs/2022arXiv221007490Y>.
13. Peng, Y., Kim, J., Feng, D. & Bi, L. Automatic Tumor Segmentation via False Positive Reduction Network for Whole-Body Multi-Modal PET/CT Images. arXiv:2209.07705 (2022).

<https://ui.adsabs.harvard.edu/abs/2022arXiv220907705P>.

14. Ma, J. & Wang, B. *nnU-Net for Automated Lesion Segmentation in Whole-body FDG-PET/CT*, <https://github.com/JunMa11/PETCTSeg/blob/main/technical_report.pdf> (2022).
15. Zhang, J., Huang, Y., Zhang, Z. & Shi, Y. Whole-Body Lesion Segmentation in 18F-FDG PET/CT. arXiv:2209.07851 (2022). <https://ui.adsabs.harvard.edu/abs/2022arXiv220907851Z>.
16. Heiliger, L. *et al.* AutoPET Challenge: Combining nn-Unet with Swin UNETR Augmented by Maximum Intensity Projection Classifier. arXiv:2209.01112 (2022). <https://ui.adsabs.harvard.edu/abs/2022arXiv220901112H>.
17. Sibille, L., Zhan, X. & Xiang, L. Whole-body tumor segmentation of 18F -FDG PET/CT using a cascaded and ensembled convolutional neural networks. arXiv:2210.08068 (2022). <https://ui.adsabs.harvard.edu/abs/2022arXiv221008068S>.
18. Bendazzoli, S. & Astaraki, M. PriorNet: lesion segmentation in PET-CT including prior tumor appearance information. arXiv:2210.02203 (2022). <https://ui.adsabs.harvard.edu/abs/2022arXiv221002203B>.
19. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* **18**, 203-211 (2021). https://doi.org:10.1038/s41592-020-01008-z
20. Erickson, N. *et al. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data*. (2020).
21. Gatidis, S., Küstner, T., Ingrisch, M., Fabritius, M. & Cyran, C. *Automated Lesion Segmentation in Whole-Body FDG- PET/CT*. (Zenodo, 2022).
22. Gatidis, S. *et al.* A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions. *Scientific Data* **9**, 601 (2022). https://doi.org:10.1038/s41597-022-01718-3
23. Sundar, L. K. S. *et al.* Fully Automated, Semantic Segmentation of Whole-Body (18)F-FDG PET/CT Images Based on Data-Centric Artificial Intelligence. *J Nucl Med* **63**, 1941-1948 (2022). https://doi.org:10.2967/jnumed.122.264063

Figures

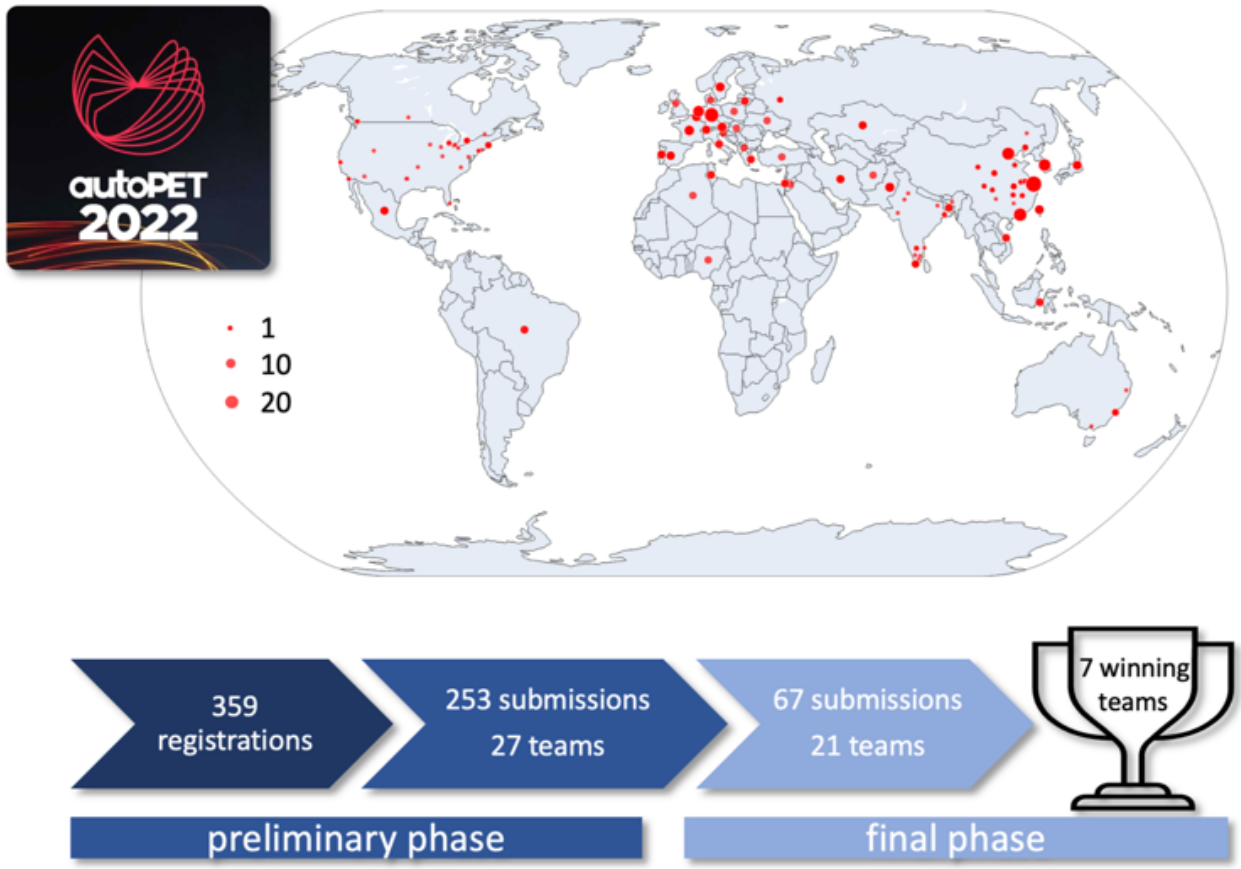

Figure 1

Challenge organization and participation

The autoPET challenge, conducted in 2022, consisted of two phases: a preliminary phase - allowing participants to perform technical validation of their algorithms on a small private test set – and a final phase where participants contributed their algorithms for final evaluation on the entire private test set. The seven best performing contributions were awarded and are described in this paper. In total, 359 teams from all continents participated in this challenge. Top: Geographic distribution of registered teams, bottom: challenge phases and participation.

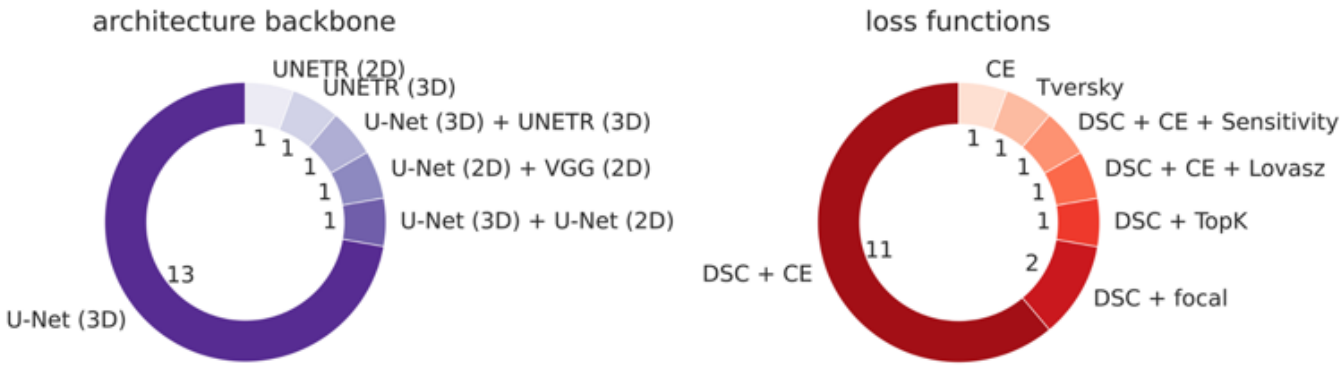

Figure 2

Overview of technical details of the final phase submissions

All participants used deep learning techniques to solve the challenge task. The majority of participating teams used a 3D U-Net architecture (left) together with a combined Dice and cross-entropy (CE) loss (right).

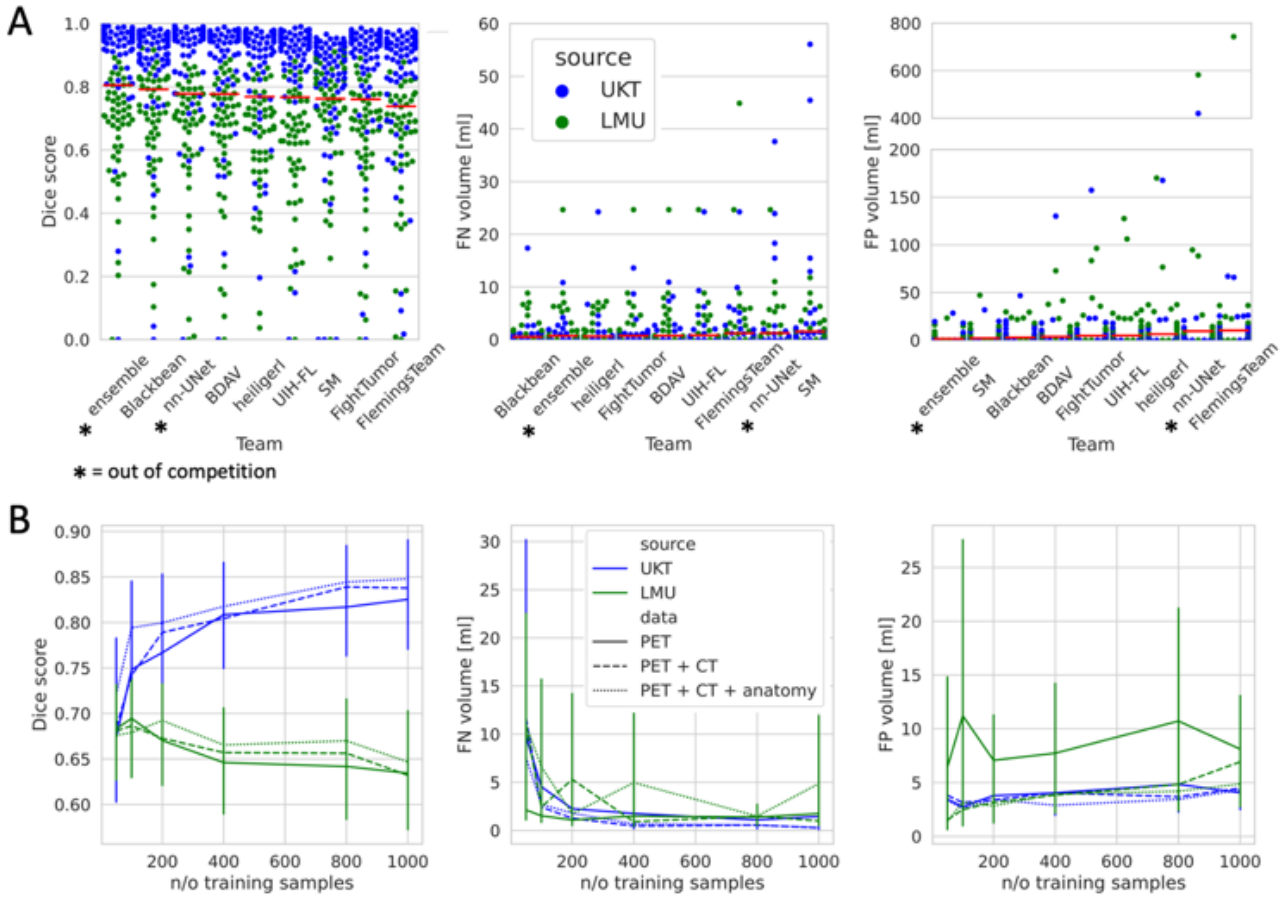

Figure 3

Overview of algorithm performance

A) Challenge results and algorithm performance in terms of Dice score, false negative (FN) volume and false positive (FP) volume. Results are ordered from best (left) to worst (right). Overall, algorithms performed better on the test data drawn from the training distribution (UKT, blue dots) compared to out-of-distribution data (LMU, green dots)

B) Impact of data source (UKT vs. LMU) and number of training samples on the performance of the baseline nn-UNet model in terms of Dice score, false negative (FN) volume and false positive (FP) volume. Overall, algorithm performance was higher on UKT test data compared to LMU test data and improved with increasing number of training samples. Notably, algorithm performance in terms of Dice scores did not improve with increasing numbers of training samples on out-of-distribution data (LMU).

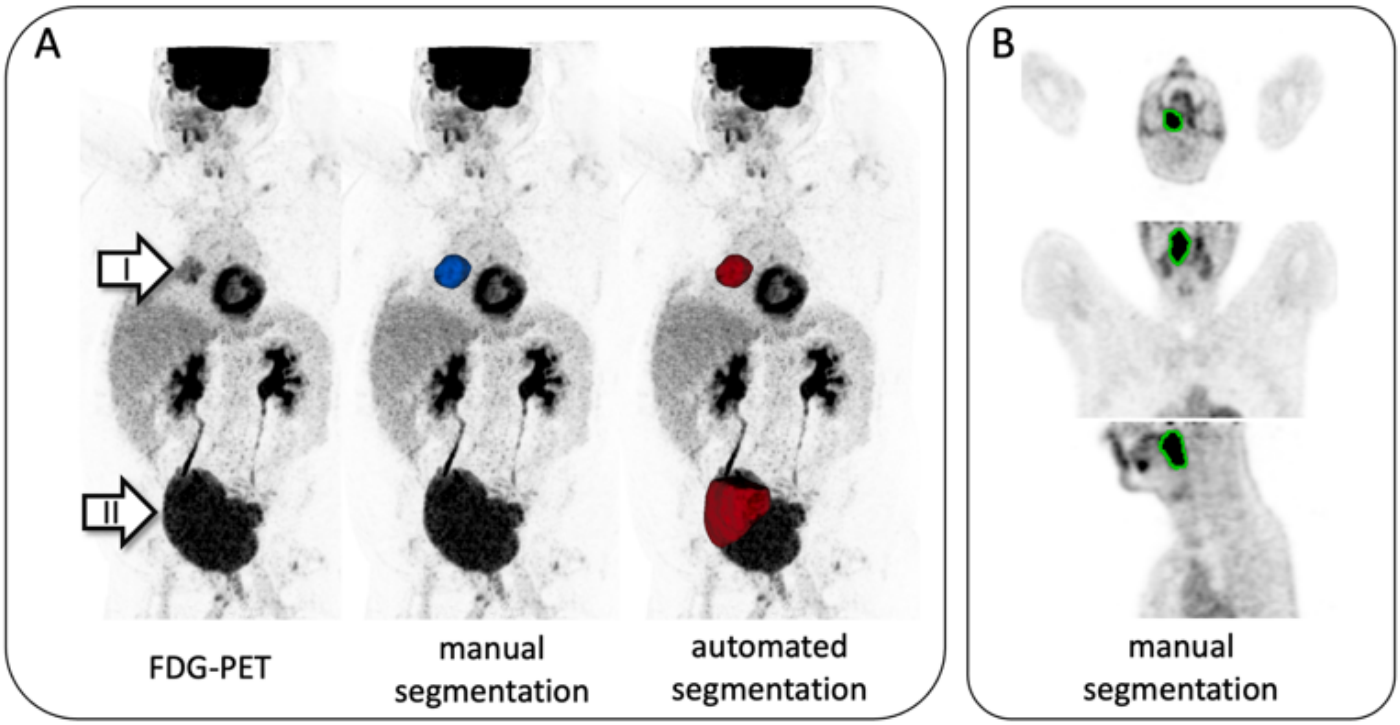

Figure 4

Qualitative examples of false positive and false negative volumes

A) Example of a large false positive volume, drawn from the UKT test data. PET scan of a patient with lung cancer (arrow I). Manual segmentation (in blue) shows the tumor lesion. Automated segmentation (red) using the baseline nn-UNet model accurately captures the tumor volume but in addition includes a large portion of the unusually large urinary bladder (arrow II). This false positive segmentation was observed in the majority of submitted algorithms.

B) Example of a false negative volume, drawn from the LMU test data. PET scan of a patient with Non-Hodgkin-Lymphoma of the right pharyngeal tonsil (manual segmentation outlined in green). This lesion was missed by 5 of the 7 best-performing contributions, probably due to its uncommon location.

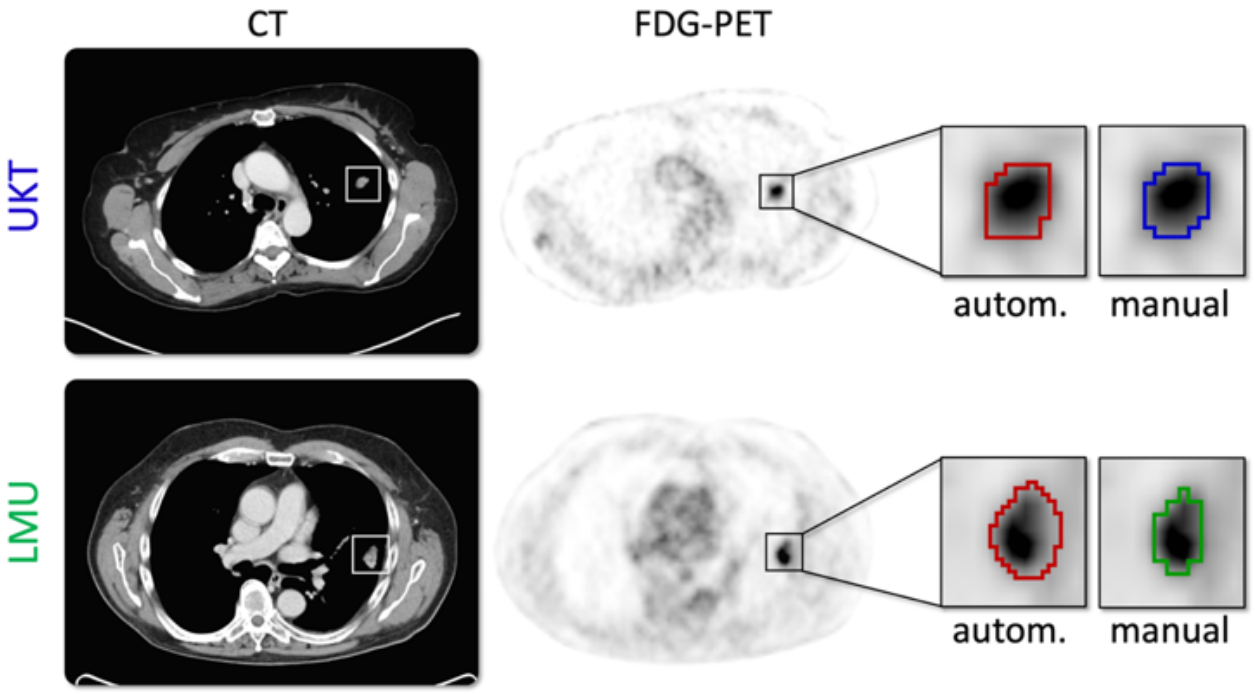

Figure 5

Typical examples of PET/CT data and segmentations from UKT (top) and LMU (bottom)

While CT image appearance is comparable between LMU and UKT data, PET scans have a lower spatial resolution on LMU data. As a result, all algorithms tended to overestimate local tumor volumes on LMU data (right column, outlined in green) compared to the manual ground truth segmentation (outlined in red). On UKT test data, automated tumor segmentations (right column, outlined in blue) showed a better agreement with manual (outlined in red) and segmentations. This is also reflected in the overall lower Dice scores of automated lesion segmentation on LMU test data compared to UKT test data.

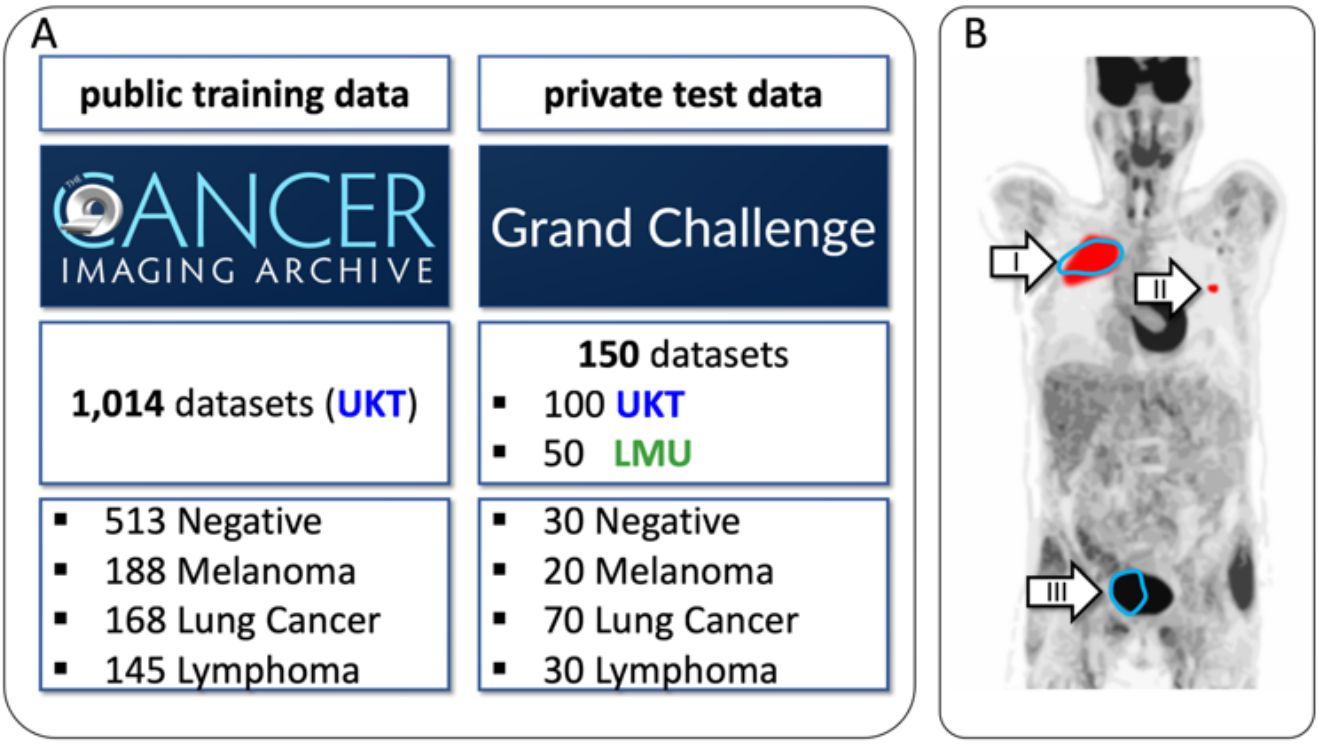

Figure 6

Challenge data and Evaluation Metrics

A) Overview of the composition of training data and test data. Training data were public and drawn from a single institution and scanner (UKT). Test data were private and drawn from two institutions: UKT (same as training data distribution) and LMU (out of training distribution).

B) Schematic illustration of the challenge metrics. I: a primary tumor lesion (red), II: a metastasis (red). Blue: Algorithm output. The Dice score provides a measure of the overall overlap between tumor lesions and algorithm segmentation. In this illustration, the lesion II is a false negative volume, as it is entirely not captured by the algorithm. Segmentation III (partial segmentation of the urinary bladder) is a false positive volume as it is not related to a tumor lesion.

Supplementary Files

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

- SupplementalFigures.docx