Automated Lesion Segmentation in Whole-Body PET/CT - Multitracer Multicenter generalization: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

Use the title to convey the essential information on the challenge mission.

Automated Lesion Segmentation in Whole-Body PET/CT - Multitracer Multicenter generalization

Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

AutoPET III

Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Positron Emission Tomography/Computed Tomography (PET/CT) has become an integral diagnostic imaging modality for various oncological indications over the past two decades. A crucial initial processing step for quantitative PET/CT analysis is segmentation of tumor lesions enabling accurate feature extraction, tumor characterization, oncologic staging, and image-based therapy response assessment. However, the growing volume of examinations, the introduction of novel PET tracers, and the increasing demand for sophisticated quantitative analyses have intensified the complexity and time requirements of these procedures.

In the first run of the autoPET challenge (AutoPET I at MICCAI 2022) we provided training and test data from University Hospital Tübingen (UKT) and LMU Hospital (LMU) as well as a baseline model for automated lesion segmentation on whole-body Fluorodeoxyglucose (FDG) PET/ CT data. The challenge results demonstrated the feasibility of accurate lesion segmentation.

In the second edition of the AutoPET challenge (AutoPET II at MICCAI 2023) we shifted our focus towards evaluating algorithmic robustness across different environments. This involved introducing a new test set comprising samples from various sites, utilizing different tracers and presenting diverse pathologies. We also relaxed constraints on using external and additional data. The challenge results revealed a substantial decline in performance, highlighting the critical need for the development of robust algorithms.

Based on the insights of the last two challenges, we propose to expand the scope of the AutoPET III challenge to the primary task of achieving multitracer multicenter generalization of automated lesion segmentation.
To this end, we provide participants access to a second, large PET/CT training dataset. This dataset introduces a new tracer, Prostate-Specific Membrane Antigen (PSMA), encompassing 597 PET/CT volumes of male patients

diagnosed with prostate carcinoma. The data was acquired at LMU with a significant domain shift from the UKT training data provided in AutoPET I and II. Algorithms will be tested on PSMA and FDG data from LMU and UKT, respectively. We will use a mixed model framework to rank valid submissions accounting for the effects of different tracers and different sites.

In addition, we will have a second award category where participants are invited to submit our baseline model trained with their creative data pipelines. This category is motivated by the observation that in AutoPET I and II, data quality and handling in pre- and post-processing posed significant bottlenecks. Due to the rarity of PET data in the medical deep learning community, there is no standardized approach to preprocess these images (normalization, augmentations, etc.). The second award category will thus additionally promote data-centric approaches to automated PET/CT lesion segmentation.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

PET/CT, PSMA, FDG, Tumor, Segmentation, Robustness

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

If the challenge is part of a workshop, please indicate the workshop.

N/A

Duration

How long does the challenge take?

Half day.

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

We expect 30 final submissions based on the previous challenges.
Final submissions: 21 (AutoPET I), 18 (AutoPET II).
Registered participants: 273 (AutoPET I), 335 (AutoPET II).

Publication and future plans

Please indicate if you plan to coordinate a publication of the challenge results.

We aim to summarize the design, proposed methods and results of the challenge in a manuscript to be submitted to a peer-reviewed scientific journal in the field of medical image analysis. To this end, we aim to invite the best performing participants to contribute by describing their methods and experiences. Furthermore, we aim to make the code of the best performing methods publicly available for the purpose of reproduction of results and further

research.

Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

For algorithm implementation and training, the participants will use their own resources. For testing as part of the challenge we aim to use resources provided on the Grand Challenge platform.

TASK 1: PET/CT lesion segmentation

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Positron Emission Tomography/Computed Tomography (PET/CT) has become an integral diagnostic imaging modality for various oncological indications over the past two decades. A crucial initial processing step for quantitative PET/CT analysis is segmentation of tumor lesions enabling accurate feature extraction, tumor characterization, oncologic staging, and image-based therapy response assessment. However, the growing volume of examinations, the introduction of novel PET tracers, and the increasing demand for sophisticated quantitative analyses have intensified the complexity and time requirements of these procedures.

In the first run of the autoPET challenge (AutoPET I at MICCAI 2022) we provided training and test data from University Hospital Tübingen (UKT) and LMU Hospital (LMU) as well as a baseline model for automated lesion segmentation on whole-body Fluorodeoxyglucose (FDG) PET/ CT data. The challenge results demonstrated the feasibility of accurate lesion segmentation.

In the second edition of the AutoPET challenge (AutoPET II at MICCAI 2023) we shifted our focus towards evaluating algorithmic robustness across different environments. This involved introducing a new test set comprising samples from various sites, utilizing different tracers and presenting diverse pathologies. We also relaxed constraints on using external and additional data. The challenge results revealed a substantial decline in performance, highlighting the critical need for the development of robust algorithms.

Based on the insights of the last two challenges, we propose to expand the scope of the AutoPET III challenge to the primary task of achieving multitracer multicenter generalization of automated lesion segmentation.
To this end, we provide participants access to a second, large PET/CT training dataset. This dataset introduces a new tracer, Prostate-Specific Membrane Antigen (PSMA), encompassing 597 PET/CT volumes of male patients diagnosed with prostate carcinoma. The data was acquired at LMU with a significant domain shift from the UKT training data provided in AutoPET I and II. Algorithms will be tested on PSMA and FDG data from LMU and UKT, respectively. We will use a mixed model framework to rank valid submissions accounting for the effects of different tracers and different sites.

In addition, we will have a second award category where participants are invited to submit our baseline model trained with their creative data pipelines. This category is motivated by the observation that in AutoPET I and II, data quality and handling in pre- and post-processing posed significant bottlenecks. Due to the rarity of PET data in the medical deep learning community, there is no standardized approach to preprocess these images (normalization, augmentations, etc.). The second award category will thus additionally promote data-centric approaches to automated PET/CT lesion segmentation.

Keywords

List the primary keywords that characterize the task.

PET/CT, PSMA, FDG, Tumor, Segmentation, Robustness

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

Michael Ingrisch, Jakob Dexl, Katharina Jeblick, Clemens Cyran, Sergios Gatidis, Thomas Kuestner

b) Provide information on the primary contact person.

Prof. Dr. Michael Ingrisch,
LMU University Hospital - Department of Radiology,
81377 München, Germany,
Michael.Ingrisch@med.uni-muenchen.de

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Repeated event as open call challenge.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

N/A

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

The challenge website will be opened on grand-challenge (similar to the last challenge: https://autopet-ii.grand-challenge.org/).

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Only publicly available data is allowed.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate but not eligible for awards and not listed in leaderboard.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

We foresee the following awards for the best submissions in two categories:
Category 1: Highest overall score
1st Place: 2,000 €
2nd Place: 1,000 €
3rd Place: 500 €

Category 2: Highest overall score for data-centric baseline model
1st Place: 1,500 €
2nd Place: 1,000 €
3rd Place: 500 €

In the event of a tie the aggregate prize allocated for the tied positions shall be combined and equitably divided among the participants who share the identical ranking.

We have applied for funding for this challenge at German cancer aid (Deutsche Krebshilfe) and other potential sponsors. Depending on the outcome of this process this award policy may be adapted.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

All submissions will be reported in the leaderboard. Participating teams can opt out of publication of their results in the leaderboard. Prize-winning methods will be announced publicly as part of a scientific session at the MICCAI annual meeting.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

The prize-winning teams will be invited to contribute to draft and submit a manuscript describing the methods and results of the challenge in a peer-reviewed journal. The first and last author of the prize-winning submissions will be also listed as authors of this planned manuscript. The participating teams may publish their own results separately after coordination to avoid significant overlap with the challenge paper.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Algorithms will be accepted as Docker containers according to the technical requirements of grand-challenge.org. Submission details will be published at the time point of challenge announcement.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

In order to enable participants to assess the technical compatibility of their to-be-submitted algorithm, we will provide information on successful algorithm deployment upon submission. To this end, the submitted algorithms will be applied to a small number of test data ensuring technical compatibility. This will allow participants to resubmit their algorithms in case of technical failure.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Challenge announcement and release of training cases: 03/2024
Registration: starting 04/2024
Submission deadline: 09/2024

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

For the already released FDG data: The ethics committee of the University Hospital Tuebingen was consulted regarding anonymized publication of training data. Ethical approval was obtained for the anonymized training database.

For the new PSMA dataset: The challenge dataset will be released in an anonymized fashion on the cancer imaging archive (TCIA), following the rigorous and well-established de-identification approach of TCIA. Our institutional data security and privacy review board has approved the de-identification approach, and our

institutional review board has waived the need for informed consent and approved the publication of the anonymized dataset (Ethics Committee, Medical Faculty, LMU Munich).

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

Code for algorithm evaluation will be available at the time point of challenge submission on: www.github.com

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

Publication of algorithm code will be a prerequisite for award eligibility. To this end code will need to be published on a publicly accessible repository of the teams choice within a week after submission deadline.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

This challenge is an initiative of the Radiology Departments of the University Hospitals of Tuebingen and LMU together with the European Society of Radiology (ESR) and the European Society for Hybrid, Molecular and Translational Imaging (ESHI-MT). We aim to seek funding for this challenge from major companies in the field of medical imaging and data analysis. So far, no concrete sponsoring has been agreed on. Test cases and labels will only be available to a limited number of colleagues involved in organizing this challenge at the participating institutions.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase
-

Research, CAD, Diagnosis

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Segmentation

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

The target cohort consists of patients undergoing FDG and PSMA PET/CT examinations in an oncologic context for diagnosis, staging, or therapy response assessment.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

The FDG cohort comprises 501 patients diagnosed with histologically proven malignant melanoma, lymphoma, or lung cancer, along with 513 negative control patients. The PSMA cohort includes pre- and/or post-therapeutic PET/CT images of male individuals with prostate carcinoma, encompassing images with (537) and without PSMA-avid tumor lesions (60). Notably, the training datasets exhibit distinct age distributions: the FDG UKT cohort spans 570 male patients (mean age: 60; std: 16) and 444 female patients (mean age: 58; std: 16), whereas the PSMA MUC cohort tends to be older, with 597 male patients (mean age: 71; std: 8). Additionally, there are variations in imaging conditions between the FDG Tübingen and PSMA Munich cohorts, particularly regarding the types and number of PET/CT scanners utilized for acquisition. The PSMA Munich dataset was acquired using three different scanner types (Siemens Biograph 64, Siemens Biograph 20 mCT and GE Discovery 690), whereas the FDG Tübingen dataset was acquired using a single scanner (Siemens Biograph mCT). Additional information about the PET/CT imaging protocols is provided in the section "Data acquisition details".

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

All PET/CT data within this challenge have been acquired on state-of-the-art PET/CT scanners using standardized protocols following international guidelines. CT as well as PET data are provided as 3D volumes consisting of stacks of axial slices.

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

No additional information will be provided regarding the image data.

b) ... to the patient in general (e.g. sex, medical history).

Age, sex, diagnosis

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Data provided as part of this challenge consists of whole-body PET/CT examinations. Usually, the scan range of these examinations extends from the skull base to the mid-thigh level. If clinically relevant, scans can be extended to cover the entire body including the entire head and legs/feet. PSMA PET/CT training volumes do not include the head.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

The target structures of the to-be-developed algorithms are FDG- and PSMA-avid malignant tumors (i.e. primary tumors and metastases).

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Aim I.) Accurate segmentation of FDG- and PSMA-avid tumor lesions in whole-body PET/CT images. The specific challenge in automated segmentation of lesions in PET/CT is to avoid false-positive segmentation of anatomical structures that have physiologically high uptake while capturing all tumor lesions. This task is particularly challenging in a multitracer setting since the physiological uptake partly differs for different tracers: e.g. brain, kidney, heart for FDG and e.g. liver, kidney, spleen, submandibular for PSMA.

Aim II.) Robust behavior of the to-be-developed algorithms with respect to moderate changes in the choice of tracer, acquisition protocol or acquisition site. This will be reflected by the test data which will be drawn partly from the same distribution as the training data and partly from a different hospital with a similar, but slightly different acquisition setup.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

PET/CT data were acquired on state-of-the-art PET/CT scanners (Siemens Biograph mCT, Siemens Biograph 64, GE Discovery 690).

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

PET/CT acquisition protocols:

Training data:
FDG dataset: Patients fasted at least 6 h prior to the injection of approximately 350 MBq 18F-FDG. Whole-body PET/CT images were acquired using a Biograph mCT PET/CT scanner (Siemens, Healthcare GmbH, Erlangen, Germany) and were initiated approximately 60 min after intravenous tracer administration. Diagnostic CT scans of the neck, thorax, abdomen and pelvis (200 reference mAs; 120 kV) were acquired 90 sec after intravenous injection of a contrast agent (90-120 ml Ultravist 370, Bayer AG) or without contrast agent (in case of existing contraindications). PET Images were reconstructed iteratively (three iterations, 21 subsets) with Gaussian

post-reconstruction smoothing (2 mm full width at half-maximum). Slice thickness on contrast-enhanced CT was 2 or 3 mm.

PSMA dataset: Examinations were acquired on different PET/CT scanners (Siemens Biograph 64, Siemens Biograph 20 mCT and GE Discovery 690). The imaging protocol mainly consisted of a diagnostic CT scan from the skull base to the mid-thigh using the following scan parameters: reference tube current exposure time product of around 140 mAs (mean); tube voltage of 100kV or 120 kV for most cases, slice thickness of 2.5 - 5 mm. Intravenous contrast enhancement was used in most studies, except for patients with contraindications.
The whole-body PSMA-PET scan was acquired on average around 71 minutes after intravenous injection of 246 MBq 18F-PSMA (mean, 369 studies) or 214 MBq 68Ga-PSMA (mean, 228 studies), respectively. The PET data was reconstructed with attenuation correction derived from corresponding CT data using standard, vendor-provided image reconstruction algorithms (e.g. an ordered-subset expectation maximization (OSEM) algorithm) with a slice thickness ranging from 3 - 5 mm (mean: 3.5 mm).

Test data: Test data will be drawn in part (50%) from the same sources and distributions as the training data. The other part will be drawn crosswise from the other center, i.e. PSMA from Tuebingen (25%) and FDG from LMU (25%). At this moment we will not disclose details of test data as we aim to avoid fine-tuning of algorithms to the test data domain. The distribution of test data will be made public after the challenge deadline.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

Data were acquired from two large university hospitals (UKT, LMU).

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Data were acquired by specialized teams consisting of Radiologists, Nuclear Medicine Physicians, and Technologists. Manual segmentation of tumor lesions was performed by two Radiologists with experience in Hybrid Imaging.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

A case (training or test case) consists of one 3D whole-body PET volume, one corresponding 3D whole-body CT volume and one 3D binary mask of manually segmented tumor lesions of the size of the PET volume. CT and PET

were acquired simultaneously on a single PET/CT scanner in one session; thus PET and CT are anatomically aligned up to minor shifts due to physiological motion. A pre-rocessing script for resampling the PET and CT to the same matrix size will be provided.

b) State the total number of training, validation and test cases.

Training cases: 1014 FDG studies and 597 PSMA studies.
Test cases: 200 (50 FDG LMU, 50 FDG Tuebingen, 50 PSMA LMU, 50 PSMA Tuebingen).

Challenge participants are free to use additional external training data or pre-trained models as long as these are publicly available. If additional labels are generated in this process, these should be made publicly available after the challenge. The use of additional training data or models should be reported.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

The number of training cases is a trade-off between the effort of manual lesion segmentation and the necessity for sufficient training data. To our knowledge this is the largest publicly available labeled whole body PET/CT data set and the first large labeled PSMA PET/CT dataset.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

All training cases are drawn from two Hospitals (University Hospital Tubingen, LMU University Hospital Munich).

Test cases are split in four subgroups: 50 FDG LMU, 50 FDG UKT, 50 PSMA LMU, 50 PSMA UKT

The rationale behind this selection of test cases is to assess the robustness of algorithms and to cross-evaluate the effects of center and tracer.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

All data were manually annotated by radiologists with experience in Hybrid Imaging. To this end, tracer-avid tumor lesions were manually segmented on the PET image data using dedicated software.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

The following annotation protocol was defined:
Step 1: Identification of tracer-avid tumor lesions by visual assessment of PET and CT information together with the clinical examination reports.
Step 2: Manual free-hand segmentation of identified lesions in axial slices.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

FDG PET/CT training and test data from UKT was annotated by a Radiologist with 10 years of experience in Hybrid Imaging and experience in machine learning research.

FDG PET/CT test data from LMU University Hospital Munich was annotated by a radiologist with 8 years of experience in hybrid imaging.

PSMA PET/CT training and test data from the LMU University Hospital Munich as well as PSMA PET/CT test data from UKT was annotated by a single reader and reviewed by a radiologist with 5 years of experience in hybrid imaging.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

N/A

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

For training and test data, original DICOM files will be anonymized and provided. In addition, scripts for conversion to the NIfTI format, for conversion of PET data from original image units (activity count) to standardized uptake values, and for resampling the PET and CT to the same matrix size will be provided.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

One relevant possible error source for image annotation is the uncertainty for whether a tracer-avid lesion is actually malignant as this cannot always be determined by PET/CT alone. For experienced readers and regarding the tumor entities chosen in this challenge, a rough estimate of false classification of lesions in this sense is about 5-10%. The second relevant possible error source is the uncertainty of defining lesion boundaries, especially on PET data that have low intrinsic resolution. The difference in segmentation volumes can range from 5-30%.

b) In an analogous manner, describe and quantify other relevant sources of error.

A further potential error source are image artifacts in PET and/or CT that may result in altered image properties.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

We will use a combination of 3 metrics reflecting the aims and specific challenges for the task of PET lesion segmentation:
1) Dice Similarity Coefficient (DSC) of segmented lesions (equal to F1 Score).
2) False positive volumes: Volume of false positive connected components that do not overlap with positives.
3) False negative volumes: Volume of false negative connected components that do not overlap with true positives.
In case of test data that do not contain positives (no lesions), only metric 2 will be used.
The post-challenge analysis of the first iteration of the AutoPET challenge (in 2022) showed that these metrics provide a comprehensive summary description of algorithm performance in terms of segmentation accuracy.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

The overarching objective of the present challenge is to develop approaches for segmentation of tracer-avid tumor lesions. We hypothesize that such automated segmentation could shift the clinical evaluation of PET/CT imaging in these patients towards a more quantitative approach, which may better support and inform clinical decision-making. While the impact of such a quantitative approach on clinical decision-making is far beyond the scope of the present challenge, we have devised our metrics for the challenge to reflect critical aspects of PET/CT tumor segmentation.

The DSC will serve as a metric for assessing the overall segmentation quality. This is feasible because, in many instances, the clinical relevant target entity is the overall lesion volume (Seifert, R. 2021).
False positive and false negative volumes are used to reflect the specific challenges of PET lesion segmentation: avoiding segmentation of physiological uptake and detecting small tumor lesions.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

1. Generalization to Specific Effects:
By treating the center and tracer as random effects, the model generalizes well to these specific effects, ensuring that the performance evaluation is not overly influenced by peculiarities associated with specific centers or tracers.

2. Post-Hoc Testing and Ranking:
The use of post-hoc Tukey tests allows for rigorous pairwise comparisons between submitted models. This statistical approach provides a systematic and objective way to determine whether observed differences in performance are statistically significant.

Overall, our mixed model approach is suitable for scenarios with diverse and nuanced factors (centers, tracers) affecting model performance. Therefore it contributes to a more robust, objective, and statistically supported evaluation of model performances.

b) Describe the method(s) used to manage submissions with missing results on test cases.

In the event of a missing test prediction, the metric score for that particular sample will be assigned the least favorable outcome.

c) Justify why the described ranking scheme(s) was/were used.

Each metric is calculated per test sample (if applicable). The model's performance is then assessed with a mixed model framework. By incorporating the center and tracer as random effects, and thereby correcting for the (fixed) effect of the actual model performance, we ensure generalization to the random effects (center and tracer). The mean performance values (= fixed effects) per model are compared post-hoc using the Tukey test, and rankings are determined based on these comparisons.
If the p-value of a pairwise comparison is not significant ($p < 0.05$), the performance of the corresponding model is considered equally good. This is applied for every metric individually. The ranks are then combined to evaluate the overall best algorithm.
For the second award category a submitted model needs to have a higher rank than the supplied baseline model.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

Assessment of Variability of Rankings
Within our approach, we assess the variability of the performance of each model by incorporating random effects such as center and tracer in the mixed model, allowing for a statistical model that explicitly captures and quantifies the sources of variability. The post-hoc Tukey test provides a formal method to determine the significance of differences in performance (ranks), offering insights into the robustness and stability of the observed model performances/rankings across various factors.

Method to assess whether Data met Assumptions
To employ our mixed model approach, the assumptions of a linear mixed model must be met, including linearity, independence, normality of residuals, and homoscedasticity. Diagnostic evaluations involve investigating (residual) plots, conducting normality checks, and potentially applying transformations to the input data.

All statistical analyses will be performed with R (https://www.r-project.org/)

b) Justify why the described statistical method(s) was/were used.

Analyzing challenge contributions with mixed effect models allows for i) accounting for random effects due to variations between centers and ii) modeling the performance of each contribution. Therefore, our mixed model analysis allows for an unbiased and objective ranking of model performance.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

Inference runtime of algorithms.

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

Gatidis S, Hepp T, Fruh M, La Fougère C, Nikolaou K, Pfannenberg C, Schölkopf B, Kustner T, Cyran C, Rubin D. A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions. Sci Data. 2022 Oct 4;9(1):601. doi: 10.1038/s41597-022-01718-3. PMID: 36195599; PMCID: PMC9532417.

Gatidis S, Früh M, Fabritius M, et al. The autoPET Challenge: Towards Fully Automated Lesion Segmentation in Oncologic PET/CT Imaging. In Review; 2023. doi:10.21203/rs.3.rs-2572595/v1

Seifert R, Kessel K, Schlack K, Weber M, Herrmann K, Spanke M, Fendler WP, Hadaschik B, Kleesiek J, Schäfers M, Weckesser M, Boegemann M, Rahbar K. PSMA PET total tumor volume predicts outcome of patients with advanced prostate cancer receiving [177Lu]Lu-PSMA-617 radioligand therapy in a bicentric analysis. Eur J Nucl Med Mol Imaging. 2021 Apr;48(4):1200-1210. doi: 10.1007/s00259-020-05040-1. Epub 2020 Sep 24. PMID: 32970216; PMCID: PMC8041668.

Pfannenberg C, Gueckel B, Wang L, Gatidis S, Olthof SC, Vach W, Reimold M, la Fougere C, Nikolaou K, Martus P. Practice-based evidence for the clinical benefit of PET/CT-results of the first oncologic PET/CT registry in Germany Eur J Nucl Med Mol Imaging. 2019 Jan;46(1):54-64. doi: 10.1007/s00259-018-4156-3. Epub 2018 Sep 29. PMID: 30269155.

https://github.com/lab-midas/autoPET

https://autopet.grand-challenge.org/

https://autopet-ii.grand-challenge.org/

Further comments

Further comments from the organizers.

N/A