scientific data

OPEN

DATA DESCRIPTOR

3D whole body preclinical micro-CT database of subcutaneous tumors in mice with annotations from 3 annotators

Malte Jensen[1], Andreas Clemmensen[1], Jacob Gorm Hansen[2], Julie van Krimpen Mortensen[1], Emil N. Christensen[1], Andreas Kjaer[1] ✉ & Rasmus Sejersten Ripa[1,3] ✉

A pivotal animal model for development of anticancer molecules is mice with subcutaneous tumors, grown by injection of xenografted tumor cells, where micro-Computed Tomography (µCT) of the mice is used to analyze the efficacy of the anticancer molecule. Manual delineation of the tumor region is necessary for the analysis, which is time-consuming and inconsistent, highlighting the need for automatic segmentation (AS) tools. This study introduces a preclinical µCT database, comprising 452 whole-body scans from 223 individual mice with subcutaneous tumors, spanning ten diverse µCT datasets conducted between 2014 and 2020 on a preclinical PET/CT scanner, making it the hitherto largest dataset of its kind. Each tumor is annotated manually by three expert annotators, allowing for robust model development. Inter-annotator agreement was analyzed, and we report an overall annotation agreement of 0.903 ± 0.046 (mean ± std) Fleiss' Kappa and a mean deviation in volume estimation of 0.015 ± 0.010 cm^3 (6.9% ± 4.7), which establishes a human baseline accuracy for delineation of subcutaneous tumors, while showing good inter-annotator agreement.

Background & Summary

One of the most commonly used imaging technologies in preclinical research is micro Computed Tomography (µCT)[1,2], with over 22.000 entries on PubMed for the keyword "micro-CT" up to this date. It offers high resolution, fast acquisition and well calibrated voxel intensities, giving detailed insights into volumes and internal structures of small animals[3]. It has high reproducibility and can be utilized both as a standalone image modality or combined with nuclear imaging such as Positron Emission Tomography (PET) or Single Photon Emission Computed Tomography (SPECT)[4].

Longitudinal studies can be performed with µCT, as the radiation dose is low. This enables monitoring of disease and treatment progression in the same animal by performing multiple scans, thus extracting more information per animal. This reduces the number of animals required to conduct studies, in accordance with the animal protection 3 R aims (Refinement, Replacement and Reduction)[5].

µCT is often performed on a large scale for preclinical research, but the resulting images require further manual analysis to be useful. The current gold-standard is to perform manual delineation of regions of interest, which is both laborious and subject to high user-dependence[6,7]. This limits the reproducibility of preclinical studies, and the time needed for manual analysis can easily exceed that of the scanning procedure itself. Hence, there is an unmet need for automatic segmentation (AS) tools to mitigate the challenges of reproducibility and time consumption in preclinical imaging studies. Automatic segmentation models are machine learning models that once trained, can take in a new image and decide what label should be assigned to each pixel in the image without any human intervention needed.

Recently, with the introduction of machine learning algorithms, repetitive tasks that require human interaction can be automated by training models on large datasets. The use of machine learning algorithms for AS offers

[1]Department of Clinical Physiology and Nuclear Medicine & Cluster for Molecular Imaging, Copenhagen University Hospital – Rigshospitalet & Department of Biomedical Sciences, University of Copenhagen, Copenhagen, Denmark. [2]Vertigo.ai, Copenhagen, Denmark. [3]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. ✉e-mail: akjaer@sund.ku.dk; rasmus.ripa@regionh.dk

Dataset	Scans, n	Mice, n	Tumors per mouse	Volume ($cm^3 \pm$ std)	Minimum (cm^3)	Maximum (cm^3)	Year of scans
Dataset 1	43	15	2	0.330 ± 0.153	0.098	0.766	2014
Dataset 2	61	32	1	0.185 ± 0.127	0.044	0.551	2014
Dataset 3	24	18	1	0.329 ± 0.217	0.071	0.793	2014
Dataset 4	40	10	1	0.094 ± 0.065	0.014	0.240	2016
Dataset 5	54	28	1	0.932 ± 0.955	0.063	3.951	2016
Dataset 6	37	18	2	0.118 ± 0.051	0.025	0.310	2017
Dataset 7	12	4	1	0.796 ± 0.431	0.388	1.528	2017
Dataset 8	91	40	1	0.466 ± 0.329	0.056	1.947	2020
Dataset 9	28	18	1	0.460 ± 0.276	0.132	1.220	2020
Dataset 10	62	40	2	0.101 ± 0.054	0.020	0.353	2020

Table 1. Details for each dataset. Std = Standard Deviation.

the prospect of improving reproducibility, consistency, and reliability in the analysis, and thus a possible solution to the aforementioned challenges in the analysis of preclinical images.

A widespread disease model in image-based preclinical research is immunosuppressed mice in which xenografted tumor cells from human cancers have been injected under the skin, which then develop into human-like subcutaneous tumors[8–11]. These models are a staple for human anti-cancer drug discovery, where the drug uptake in the tumor can be measured as well as the tumor growth rate and tumor metabolism. This model can further be utilized for personalized medicine by xenografting individual human tumor biopsies to assess the sensitivity to different anti-cancer agents in a patient-to-patient approach.

While several approaches to AS on medical images exist, there are no public AS models or datasets for subcutaneous tumors in neither µCT scans[12] or Magnetic Resonance Imaging (MRI) scans[13]. Research on AS for other types of tumors has been done for µCT[14,15], as well as for MRI[16–21], and cryogenic-imaging[22]. However, these were generally performed on small datasets, which limit their usefulness as a general tool. Also, they have not made their models publicly available.

Classically, the approaches to AS have often been atlas-based algorithms, where one or multiple anatomical atlases guide the AS[23,24], or filter-based, where a large set of filters are used to extract features for a machine learning algorithm[25]. However, subcutaneous tumors differ significantly in anatomical placement and morphology between each mouse, which makes them less suitable for atlas-based algorithms. The texture of the tumor and surrounding soft-tissue is quite similar, which also makes texture-based methods less suitable. Deep learning models excel in learning complex interactions between morphology and texture, but require large amounts of high-quality data to be trained successfully[26]. Our dataset is aimed at filling this data gap and enable deep learning models to be trained for this segmentation task.

We provide a preclinical µCT whole-body database of mice with subcutaneous tumors, publicly available. It consists of 452 whole-body µCT scans from 223 individual mice, retrospectively collected from ten different datasets at our institution spanning the years 2014 to 2020. All scans are annotated by three trained annotators, which gives our dataset the size and diversity needed for developing robust AS algorithms, as well as providing a human baseline for inter-annotator agreement.

The aim is that our database will serve as a resource to train and validate machine learning algorithms for AS, thus facilitating the development of fast, robust, and reproducible analysis tools for subcutaneous tumor models.

Methods

Datasets. Ten µCT datasets from 2014 to 2020 were collected (Table 1 and Figure 1). All animal experiments were approved by the Danish Animal Experiments Inspectorate (permit number 2012-15-2934-00064 and 2016-15-0201-00920). The animals were housed in the core animal facilities at the University of Copenhagen, Denmark, where they were exposed to a 12:12 hours light/dark cycle, with a temperature of 21 ± 2 °C, and access to water and rodent food ad libitum. The animals were acclimatized for at least one week before being included in the experiments. The included µCT scans have not previously been published but were only used to anatomically guide the extraction of values from corresponding PET images.

The 10 datasets collectively contain 452 µCT scans of 223 individual mice. The mice were scanned longitudinally at different time intervals on a preclinical µCT/PET scanner (Inveon, Siemens, USA). All scans were performed on athymic nude mice with human xenografted tumor cells, which had been allowed to develop into subcutaneous tumors prior to performing the scans. In 3 of the 10 datasets, each animal had two tumors: one on each flank. In the remainder of the 7 datasets, each animal had one tumor on the flank (Table 1 and Figure 1). In Dataset 8, the mice had the tumor inoculated behind the front legs instead of the flank. Mice with external necrosis on tumors and a total tumor burden of over 2000 µL were euthanized due to ethical concerns, and a typical humane endpoint would be 1500 µL. In Dataset 8 and 10, the mice were scanned in a small animal bed, and the remainder of the mice were scanned laying freely on the bed, which reflects different real world scanning scenarios.

The mice were all aged from 6–8 weeks at the time of enrolment into the experiments. During the µCT scanning procedure, the mice were anesthetized with a continuous flow of 1–2% Sevoflurane, while being placed on a heated bed. All scans were reconstructed using either filtered back projection or the Feldkamp Cone Beam algorithm in the vendor-supplied software (Inveon Acquisition Workplace, Siemens, USA) with a voxel size of

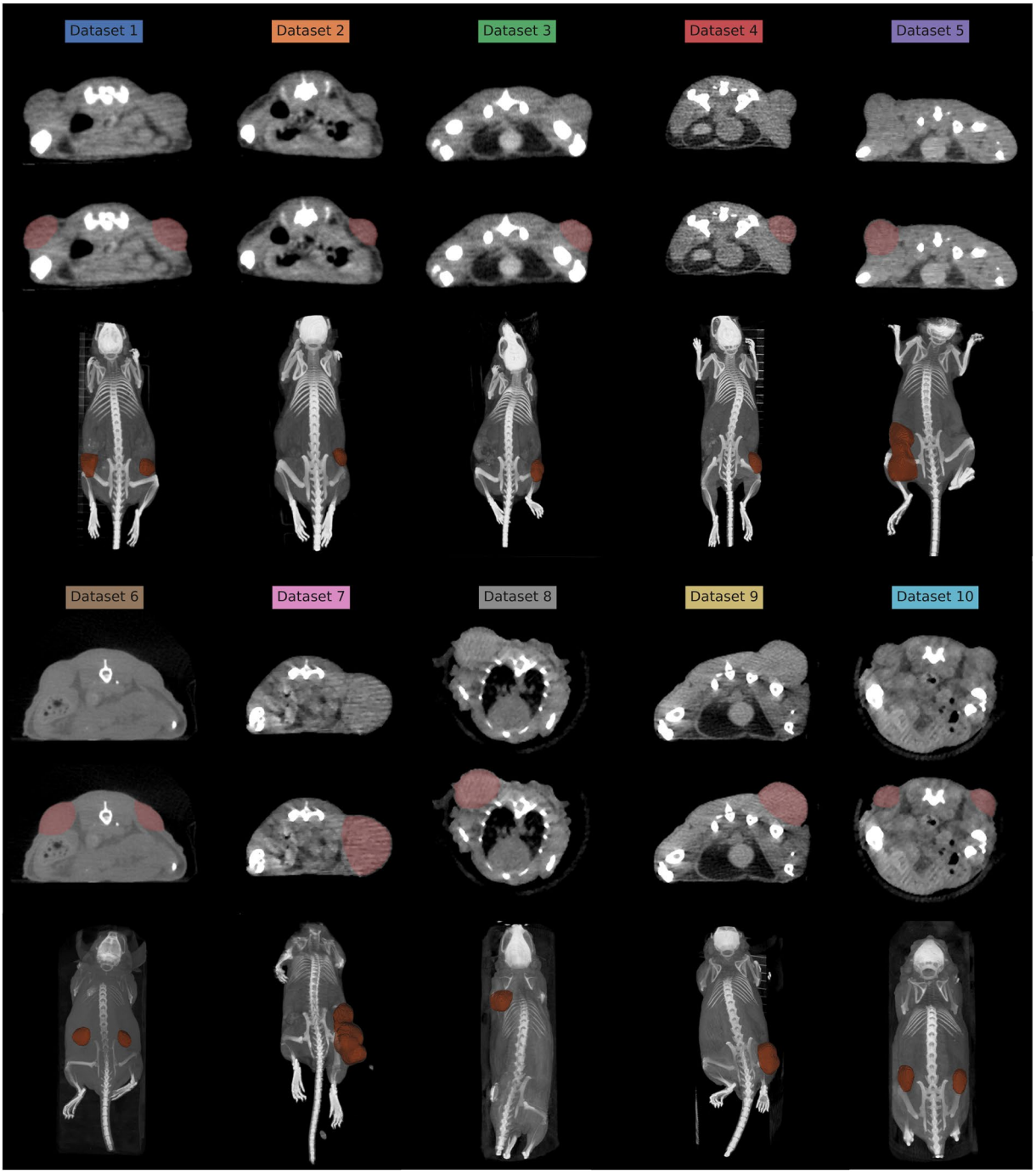

Fig. 1 Example of a µCT scan for each of the 10 datasets, with an axial slice containing the tumor with and without the tumor mask overlayed in red, and a 3D Maximum Intensity Projection of the entire scan shown below the axial slices.

$0.210 \times 0.210 \times 0.210$ mm and no spacing between slices. All scans were acquired at 500 µA, while the voltage and exposure times differed between datasets. All details can be seen in Table 2.

Data preprocessing and annotation. Each µCT scan was performed with either two or four mice in the scanner at the same time, with each mouse being placed in a small animal bed. The µCT scans were preprocessed by cropping out the mice into 192×192 pixels in the x- and y-axis, while the full length along the z-axis was kept, and then clipping the dynamic range between −400 and 1,000 Hounsfield Units. Cropping out each mouse eases the process of training machine learning algorithms as well as reducing the space needed for storage, since the mice would be surrounded by air in the field of view of the µCT scan, which the cropping process would remove the majority of.

Dataset	Voltage (kVp)	Exposure time (ms)	Scan time points	Scans *not* present at time points
Dataset 1	80	270	0 h, 3 h, 22.5 h	M11: 22.5 h; M12: 22.5 h
Dataset 2	70	350	0 h, 8d	M13: 8d; M23: 8d
Dataset 3	70	350	0 h, 24 h	M07-M18: 0 h
Dataset 4	70	280	0 h, 2 h, 6 h, 16 h	
Dataset 5	65	430	0 h, 24 h	M15: 0 h; M16: 0 h
Dataset 6	70	380	0 h, 7d, 13d, 20d	M01: 20d; M02: 20d, M03: 13d, 20d; M04: 13d, 20d; M05: 7d, 13d, 20d; M06: 20d; M07: 0d, 13d, 20d; M08: 7d; M10: 0d; M12: 13d; M11: 0d, 13d, 20d; M13: 7d, 13d, 20d; M14: 0d, 7d, 13d; M15: 13d; M16: 13d, 20d; M17: 7d, 13d, 20d; M18: 7d, 13d, 20d
Dataset 7	65	430	0 h, 3 h, 24 h	
Dataset 8	65	430	0 h, 6d, 8d	M01: 8d; M02: 8d; M06: 8d; M08: 8d; M09: 6d; M10: 6d; M12: 8d; M14: 8d; M17: 0d; M18: 0d, 8d; M19: 0d, 8d; M20: 0d; M25: 6d, 8d; M26: 6d, 8d; M27: 8d; M28: 8d; M29: 8d; M30: 8d; M31: 8d; M32: 8d; M34: 8d; M36: 8d; M39: 8d; M40: 0d, 8d;
Dataset 9	65	450	0 h, 18 h	M01: 18 h; M02: 18 h; M09: 18 h; M10: 18 h; M15: 0 h; M16: 0 h; M17: 18 h; M18: 18 h
Dataset 10	65	410	0 h, 3d	M01: 0d; M02: 13d; M11: 13d; M15: 13d; M18: 13d; M19: 13d; M20: 13d; M24: 13d; M27: 13d; M28: 0d; M29: 0d; M31: 0d; M32: 13d; M33: 13d; M39: 13d; M40: 13d

Table 2. Detailed scanning parameters for each dataset. All datasets had a voxel size of $0.210 \times 0.210 \times 0.210$ mm with no spacing between the slices.

After the µCT scans were preprocessed, all tumors were then manually labeled by three independent annotators, using the Napari Viewer[27] in Python 3.8 (Python Software Foundation, Delaware, USA). The tumors were annotated by drawing on every 5th axial slice and then using linear interpolation to form the 3D annotation of the tumor. Annotations touching either air or bones were automatically removed by thresholding to speed up the annotation process, followed by inspection and potential correction by the annotator if needed. A threshold of under −300 HU for air and over 500 HU for bones was used. If any central necrosis was present in the tumor, it was included in the annotation mask, in accordance with the RECIST guidelines[28], to ensure clinical relevance and translatability. All annotators were blinded from the dataset number and scan time of the mice during annotation to avoid biasing the delineation of the tumors.

Annotation metrics & evaluation. We used the following metrics to evaluate the annotations, which were all performed over the tumor in 3D (i.e. not slice-wise). The inter-annotator agreement was evaluated by calculating the Sørensen-Dice coefficient between annotators[29]. In our case, it was used to compare the agreement between the three pairs of annotators (A vs B, A vs C, and B vs C). The Sørensen-Dice coefficient was calculated with the following formula:

$$SD = \frac{2\,|X \cap Y|}{|X| + |Y|}$$

Where X and Y represent the set of segmented voxels by two different annotators. The Sørensen-Dice coefficient varies between 0 and 1, where a score of 1 denotes a perfect overlap between the segmentations and a score of 0 denotes no overlap.

To assess the overall agreement between the three annotators, we used Fleiss' Kappa[30]. This similarity coefficient is related to Cohen's Kappa[31] but extends to multiple annotators. In brief, Fleiss' Kappa is calculated by the following formula:

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e}$$

The denominator $1 - \overline{P}_e$ designates the degree of agreement, which is attainable above chance, while the numerator $\overline{P} - \overline{P}_e$ designates the degree of agreement that was actually achieved above chance. If the annotators are in complete agreement, then $\kappa = 1$, while if there is no agreement between the annotators above what would be expected by chance, then $\kappa \leq 0$. If the agreement is exactly the same as what is expected by chance, then $\kappa = 0$.

The agreement of volume estimation across the three annotators was estimated as the difference in estimated volume between the pairs of annotators. If two tumors were present in the mouse, their volumes were calculated individually. For comparing the agreement of all annotators on volume estimation, the Root Mean Squared Error (RMSE) between the volume estimated from each annotator and the mean of the volume from all three annotators was used. The RMSE indicates the average volume deviation of each annotator from the mean volume estimated from all annotators. The RMSE was used rather than the mean difference of all pairs, as this would trivially be zero, given the annotators were subtracted in the right order, and hence not yield any information.

The annotation results are presented in Table 3, and a detailed evaluation on a dataset-level can be seen in Fig. 2 and in Table 4.

Annotator	Sørensen-Dice coefficient	Mean volume difference (cm^3)	Mean volume difference (%)
A vs. B	0.900 ± 0.056	-0.003 ± 0.030	$-0.8\% \pm 16.5$
A vs. C	0.907 ± 0.048	0.006 ± 0.029	$2.2\% \pm 14.7$
B vs. C	0.902 ± 0.046	0.009 ± 0.030	$3.0\% \pm 15.1$
	Fleiss' Kappa	**RMSE (cm^3)**	**RMSE (%)**
All	0.903 ± 0.046	0.015 ± 0.010	6.9 ± 4.7

Table 3. Comparison between pairs of annotators and all annotators across all datasets. The metrics were calculated over the 3D tumor volume. RMSE was performed between the volume of each annotator and the mean volume across all annotators.

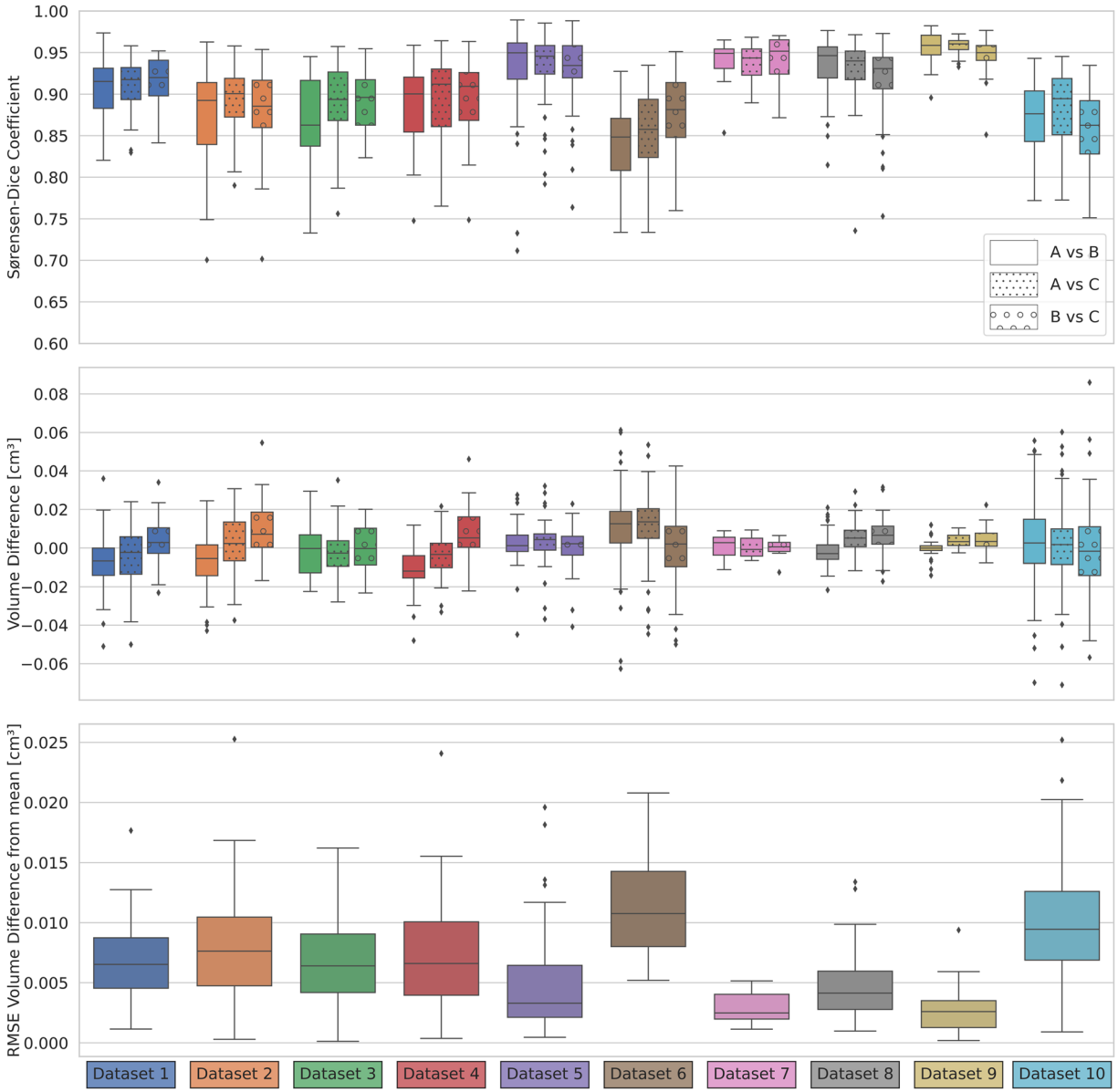

Fig. 2 Sørensen-Dice coefficient across annotators on each dataset and difference in volume. Each dataset is color-coded, while the annotator pairs are indicated by the hatching: Annotator A vs. B, A vs. C and B vs. C is shown. The middle line is the median, box ends are quartiles, whiskers are 1.5 interquartile range and dots are outliers outside the 1.5 interquartile range. (**a**) depicts the Sørensen-Dice coefficient, (**b**) depicts the difference in volume estimation and (**b**) depicts the root mean squared error between the mean tumor volume estimated from all 3 annotators, and the volume each annotator has estimated. The metrics were calculated over the 3D tumor volumes.

Dataset	A vs. B	A vs. C	B vs. C	Fleiss' Kappa
Dataset 1	0.908 ± 0.035	0.911 ± 0.031	0.915 ± 0.028	0.911 ± 0.024
Dataset 2	0.874 ± 0.055	0.891 ± 0.038	0.881 ± 0.048	0.882 ± 0.040
Dataset 3	0.870 ± 0.056	0.884 ± 0.053	0.889 ± 0.038	0.881 ± 0.042
Dataset 4	0.887 ± 0.049	0.895 ± 0.050	0.900 ± 0.043	0.894 ± 0.042
Dataset 5	0.931 ± 0.053	0.933 ± 0.042	0.928 ± 0.044	0.931 ± 0.041
Dataset 6	0.841 ± 0.047	0.853 ± 0.048	0.872 ± 0.048	0.855 ± 0.041
Dataset 7	0.938 ± 0.029	0.938 ± 0.024	0.941 ± 0.031	0.939 ± 0.026
Dataset 8	0.934 ± 0.031	0.932 ± 0.031	0.921 ± 0.036	0.929 ± 0.029
Dataset 9	0.956 ± 0.019	0.958 ± 0.010	0.946 ± 0.023	0.953 ± 0.015
Dataset 10	0.868 ± 0.046	0.883 ± 0.042	0.858 ± 0.049	0.869 ± 0.040

Table 4. Sørensen-Dice coefficient and Fleiss' Kappa on a dataset level (mean ± std) for annotator A, B and C. The metrics were calculated over the 3D tumor volume.

```
Root
|-- Dataset 1
|-- Dataset 2
    |-- M01_M02_8d
        |-- M01_8d
        |-- M02_8d
            |-- Annotator_A_M02_8d.nii.gz
            |-- Annotator_B_M02_8d.nii.gz
            |-- Annotator_C_M02_8d.nii.gz
            |-- CT_M02_8d.nii.gz
|-- Dataset 3
...
```

Fig. 3 Overview of the folder structure for the datasets. MXX is the mouse number and Xh or Xd is the hours or days since the first scan of the mouse, respectively.

Data Records

The dataset is available at the University of Copenhagen Electronic Research Data Archive[32]. The data are organized into folders for each dataset, called Dataset 1 to 10. Each dataset folder contains subfolders with mice that were scanned together (either two or four in the same scan). Each of these folders again contains subfolders with the cropped out µCT scan for each mouse, as well as the annotations from each of the 3 annotators. The mice were named MXX, where XX is the number of each mouse in the dataset. Mouse numbers will occur multiple times if the same mouse was scanned at several time points in a dataset. The scan time points appear in the names as Xh or Xd, where X is the number of hours or days since the first scan of the mouse, respectively. An overview of the folder structure can be seen in Fig. 3. The data are saved in compressed Neuroimaging Informatics Technology Initiative (NIfTI) format[33], which is compatible with most platforms for medical images. Detailed descriptions of the datasets can be found in Table 2. The xenograft tumor cell line information was not available or was of proprietary nature, and therefore, it is not included in this dataset.

Technical Validation

The presented dataset offers a basis for both development and evaluation of AS algorithms. It further establishes a baseline for human inter-annotator agreement. The overall agreement of the annotators was 0.903 Fleiss' Kappa, and the Sørensen-Dice coefficient between pairs of annotators was around ~0.90 (Table 3). The annotator agreement was slightly higher for datasets 5, 7, 8, and 9 compared to the rest of the datasets (Fig. 2 and Table 4), which was likely due to the tumors being larger in these datasets. The degree of agreement was similar to what other studies with manual segmentation of CT images have reported[12,34–39]. For example, in Rosenhain *et al.*[12] the inter-annotator agreement was 0.810 Sørensen-Dice coefficient for tumors in contrast-enhanced µCT scans, and at most 0.879 Sørensen-Dice coefficient for the organs. As an clinical example, Patil *et al.*[38] obtained a Sørensen-Dice coefficient of 0.89-0.90 for lung tumors on human CT scans. Our finding of around 0.90 Sørensen-Dice coefficient between the annotators was hence reasonable compared to similar datasets.

For the estimation of the tumor volume, each annotator pair had a mean disagreement close to zero mL across all datasets, with a standard deviation of about 0.030 mL. The RMSE from the mean volume was 0.015 mL across all datasets. In Dataset 4, 6, 9, and 10, the annotators had slightly lower variance on the agreement in volume, compared to the rest of the datasets (Fig. 3 and Table 3). This was most likely due to the image quality being slightly higher for these datasets. We note that these results are specific to human xenografts, and that other tumor models such as syngeneic models could elicit different results.

Usage Notes

All interested researchers are highly encouraged to download the 3D μCT dataset and use it for their own experiments and model development. It can be used to train AS algorithms and evaluate their accuracy against human annotators or be used as an external evaluation dataset for AS algorithms, which are trained on a different dataset. Since the dataset is annotated by three individual researchers, all annotations can be utilized in the training of the AS algorithms to yield more general and de-biased models.

When evaluating AS algorithms on our dataset, we suggest that users test and report their performance on each individual annotator's annotations, as well as the mean performance across all annotators. We have further supplied annotations that are merged from the three annotators by the STAPLE[40] algorithm, which can be additionally used to report performance of an AI model.

Having multiple annotations can further be used to develop and evaluate uncertainty quantification algorithms, as the uncertainty for each scan can be calculated through the three different annotations[41]. The dataset can further be used in training deep learning algorithms on other tasks than subcutaneous tumor segmentations, e.g. annotating new anatomical structures or for self-supervised pretraining. The NIfTI format ensures that the scans are compatible with a broad array of commercial and non-commercial software.

Code availability

No custom code was used for this paper.

Received: 17 May 2024; Accepted: 21 August 2024;
Published online: 19 September 2024

References

1. Rosenthal, N. & Brown, S. The mouse ascending: Perspectives for human-disease models. *Nature Cell Biology* **9**, 993–999 (2007).
2. Osuchowski, M. F. *et al*. Abandon the mouse research ship? Not just yet! *Shock* **41**, 463–475 (2014).
3. Perrin, S. Make mouse studies work. *Nature* **507**, 423–425 (2014).
4. James, M. L. & Gambhir, S. S. A molecular imaging primer: Modalities, imaging agents, and applications. *Physiological Reviews* **92**, 897–965 (2012).
5. Burch, R. & Russell, W. The Principles of Humane Experimental Technique by W.M.S. Russell and R.L. Burch. *John Hopkins Bloomberg School of Public Health* (1959).
6. Lappas, G. *et al*. Inter-observer variability of organ contouring for preclinical studies with cone beam Computed Tomography imaging. *Physics and Imaging in Radiation Oncology* **21**, 11–17 (2022).
7. Hecksel, C. W. *et al*. Quantifying Variability of Manual Annotation in Cryo-Electron Tomograms. *Microscopy and Microanalysis* **22**, 487–496 (2016).
8. Morton, C. L. & Houghton, P. J. Establishment of human tumor xenografts in immunodeficient mice. *Nature Protocols* **2**, 247–250 (2007).
9. Xu, C., Li, X., Liu, P., Li, M. & Luo, F. Patient-derived xenograft mouse models: A high fidelity tool for individualized medicine (review). *Oncology Letters* **17**, 3–10 (2019).
10. Richmond, A. & Yingjun, S. Mouse xenograft models vs GEM models for human cancer therapeutics. *DMM Disease Models and Mechanisms* **1**, 78–82 (2008).
11. Zitvogel, L., Pitt, J. M., Daillère, R., Smyth, M. J. & Kroemer, G. Mouse models in oncoimmunology. *Nature Reviews Cancer* **16**, 759–773 (2016).
12. Rosenhain, S. *et al*. A preclinical micro-computed tomography database including 3D whole body organ segmentations. *Scientific Data* **5**, 1–9 (2018).
13. Hectors, S. J. C. G., Jacobs, I., Strijkers, G. J. & Nicolay, K. Automatic segmentation of subcutaneous mouse tumors by multiparametric MR analysis based on endogenous contrast. *Magnetic Resonance Materials in Physics, Biology and Medicine* **28**, 363–375 (2015).
14. van de Worp, W. R. P. H. *et al*. Deep learning based automated orthotopic lung tumor segmentation in whole-body mouse ct-scans. *Cancers* **13** (2021).
15. Mateos-Pérez, J. M., Soto-Montenegro, M. L., Peña-Zalbidea, S., Desco, M. & Vaquero, J. J. Functional segmentation of dynamic PET studies: Open source implementation and validation of a leader-follower-based algorithm. *Computers in Biology and Medicine* **69**, 181–188 (2016).
16. Rallapalli, H. *et al*. MEMRI-based imaging pipeline for guiding preclinical studies in mouse models of sporadic medulloblastoma. *Magnetic Resonance in Medicine* **83**, 214–227 (2020).
17. Tidwell, V. K., Garbow, J. R., Krupnick, A. S., Engelbach, J. A. & Nehorai, A. Quantitative analysis of tumor burden in mouse lung via MRI. *Magnetic Resonance in Medicine* **67**, 572–579 (2012).
18. Holbrook, M. D. *et al*. Mri-based deep learning segmentation and radiomics of sarcoma in mice. *Tomography* **6**, 23–33 (2020).
19. Akselrod-Ballin, A. *et al*. Multimodal Correlative Preclinical Whole Body Imaging and Segmentation. *Scientific Reports* **6** (2016).
20. Gomes, A. L. *et al*. Cardio-Respiratory synchronized bSSFP MRI for high throughput *in vivo* lung tumour quantification. *PLoS ONE* **14** (2019).
21. Lam, W. W. *et al*. An Automated Segmentation Pipeline for Intratumoural Regions in Animal Xenografts Using Machine Learning and Saturation Transfer MRI. *Scientific Reports* **10** (2020).
22. Liu, Y. *et al*. Quantitative analysis of metastatic breast cancer in mice using deep learning on cryo-image data. *Scientific Reports* **11** (2021).
23. BachCuadra, M., Duay, V. & Thiran, J. P. H. Atlas-based segmentation. in *Handbook of Biomedical Imaging: Methodologies and Clinical Research* 221–244, https://doi.org/10.1007/978-0-387-09749-7_12. (Springer US, 2015).
24. Matula, J. *et al*. X-ray microtomography-based atlas of mouse cranial development. *GigaScience* **10** (2021).
25. Randen, T. & Husøy, J. H. Texture segmentation using filters with optimized energy separation. *IEEE Transactions on Image Processing* **8**, 571–582 (1999).
26. Wang, R. *et al*. Medical image segmentation using deep learning: A survey. *IET Image Processing* **16**, 1243–1267 (2022).
27. Sofroniew, N. *et al*. napari: a multi-dimensional image viewer for Python. https://doi.org/10.5281/ZENODO.7276432 (2022).
28. Eisenhauer, E. A. *et al*. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer* **45**, 228–247 (2009).
29. Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **26**, 297–302 (1945).
30. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**, 378–382 (1971).
31. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* **20**, 37–46 (1960).

32. Jensen, E. K. *et al.* 3D whole body preclinical microCT database of subcutaneous tumors in mice with annotations from 3 annotators. 2.2 gb University of Copenhagen https://doi.org/10.17894/UCPH.F7BCF864-BE18-4A16-95AD-6F22DEDB4265 (2024).
33. NIfTI: — Neuroimaging Informatics Technology Initiative. https://nifti.nimh.nih.gov/.
34. Rudd, J. H. F. *et al.* Atherosclerosis inflammation imaging with 18F-FDG PET: Carotid, iliac, and femoral uptake reproducibility, quantification methods, and recommendations. *Journal of Nuclear Medicine* **49**, 871–878 (2008).
35. Büyükdereli, G., Güler, M. & Şeydaoğlu, G. Interobserver and intraobserver variability among measurements of FDG PET/CT parameters in pulmonary tumors. *Balkan Medical Journal* **33**, 308–315 (2016).
36. Buijsen, J. *et al.* FDG-PET-CT reduces the interobserver variability in rectal tumor delineation. *Radiotherapy and Oncology* **102**, 371–376 (2012).
37. Meyers, N. *et al.* Inter-observer variability of 90Y PET/CT dosimetry in hepatocellular carcinoma after glass microspheres transarterial radioembolization. *EJNMMI Physics* **7** (2020).
38. Patil, M. B. *et al.* Inter-observer variability in the delineation of gross tumour volume GTV using PETCT in early stage non small cell lung cancer NSCLC. *Annals of Oncology* **28**, x121 (2017).
39. Jacene, H. A. *et al.* Assessment of interobserver reproducibility in quantitative 18F-FDG PET and CT measurements of tumor response to therapy. *Journal of Nuclear Medicine* **50**, 1760–1769 (2009).
40. Warfield, S. K., Zou, K. H. & Wells, W. M. Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation. *IEEE Trans Med Imaging* **23**, 903–921 (2004).
41. Kohl, S. A. A. *et al.* A Probabilistic U-Net for Segmentation of Ambiguous Images. (2018).

Author contributions

M.J., A.C., J.G.H., R.S.R. and A.K. conceived the project, and M.J. was the main responsible for the gathering, curation and labelling of the data for this paper, and performed the analysis of the data, with A.C. helping on finding the μCT datasets. A.C., J.G.H., R.S.R. and A.K. provided expert guidance and crucial ideas for the creation of the dataset and in performing the analysis. M.J., J.V.K.M. and E.N.C. provided their specialist knowledge to label the data. M.J. wrote the manuscript, and all authors have contributed with several rounds of revisions on the manuscript. A.K. provided the funding for this project.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.K. or R.S.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.