Scientific Data

https://doi.org/10.1038/s41597-025-06518-z

Article in Press

Clinically validated dataset of 435 human colons segmented from CT colonography

Received: 1 August 2025

Accepted: 19 December 2025

Cite this article as: Finocchiaro, M., Stern, R., Vilhelmsborg, R. *et al.* Clinically validated dataset of 435 human colons segmented from CT colonography. *Sci Data* (2026). https://doi.org/10.1038/s41597-025-06518-z

Martina Finocchiaro, Ronja Stern, Rikke Vilhelmsborg, Abraham George Smith, Jens Petersen, Kristoffer Cold, Lars Konge, Kenny Erleben & Melanie Ganz

SCIENTIFIC DATA

CONFIDENTIAL
COPY OF SUBMISSION FOR PEER REVIEW ONLY

Tracking no: SDATA-25-04311A

Clinically validated dataset of 435 human colons segmented from CT colonography

Authors: Martina Finocchiaro (University of Copenhagen), Ronja Stern (University of Copenhagen), Rikke Vilhelmsborg (Bispebjerg Hospital), Abraham Smith (University of Copenhagen), Jens Petersen (University of Copenhagen - Department of Computer Science DIKU), Kristoffer Mazanti Cold (Rigshospitalet), Lars Konge (Rigshospitalet), Kenny Erleben (University of Copenhagen), and Melanie Ganz (Copenhagen University Hospital, Rigshospitalet)

Abstract:
High-quality segmentation datasets are essential for advancing AI applications in medical imaging. However, it is challenging to generate such datasets for highly variable and complex organs selike the colon. We introduce a dataset of 435 human colons, segmented from Computed Tomography Colonography (CTC) obtained from the publicly available The Cancer Imaging Archive (TCIA). Each scan includes a mask of the whole colon, including collapsed segments and the fluid, and a mask of only the gas-filled parts of the colon. The colon segmentation accuracy has been clinically validated by an expert abdominal radiologist. This is the first open-access dataset of segmented colons derived from CTC. This resource enables population-scale radiologic studies, supports the development of AI-based image analysis tools, and facilitates the creation of anatomically accurate digital models and simulators, both virtual and physical.

Datasets:

Repository Name	Dataset Title	Accession Number or DOI	URL to data record	Private reviewer access URL/code
Open Science Framework (OSF)	HQColon Dataset: High-Resolution Human Colon Segmentation Dataset	10.17605/OSF.IO/8TKPM	https://osf.io/8tkpm/	

Clinically validated dataset of 435 human colons segmented from CT colonography

Martina Finocchiaro[1,*], Ronja Stern[1], Rikke Vilhelmsborg[2], Abraham George Smith[1], Jens Petersen[1], Kristoffer Cold[3], Lars Konge[3,4], Kenny Erleben[1], and Melanie Ganz[1,5]

[1]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

[2]Department of Radiology, Bispebjerg Hospital, Copenhagen, Denmark

[3]Copenhagen Academy for Medical Education and Simulation. Center for Human Resources and Education, The Capital Region of Denmark

[4]Department of Clinical Medicine, University of Copenhagen, Denmark

[5]Neurobiology Research Unit, Copenhagen University Hospital, Copenhagen, Denmark

[*]Corresponding author: martina.finocchiaro.mf@gmail.com

November 2025

Abstract

High-quality segmentation datasets are essential for advancing AI applications in medical imaging. However, it is challenging to generate such datasets for highly variable and complex organs selike the colon. We introduce a dataset of 435 human colons, segmented from Computed Tomography Colonography (CTC) obtained from the publicly available The Cancer Imaging Archive (TCIA). Each scan includes a mask of the whole colon, including collapsed segments and the fluid, and a mask of only the gas-filled parts of the colon. The colon segmentation accuracy has been clinically validated by an expert abdominal radiologist. This is the first open-access dataset of segmented colons derived from CTC. This resource enables population-scale radiologic studies, supports the development of AI-based image analysis tools, and facilitates the creation of anatomically accurate digital models and simulators, both virtual and physical.

Background & Summary

In the era of AI-driven healthcare, high-quality medical image segmentation datasets are essential for numerous applications. A large dataset of a segmented organ, such as the colon, is a valuable resource for advancing AI in medical imaging [1]. It supports the training of deep learning models for automated segmentation [2], facilitates large-scale population studies, and captures the anatomical variability essential for developing robust and generalizable tools [3]. Such datasets are also essential for objectively validating models and benchmarking their performance, as they provide a standardized basis for fair comparison between different a lgorithms. F urthermore, t hey c ontribute to the creation of digital organ representations, or "digital twins", which require extensive anatomical data [4]. Realistic anatomical models derived from these datasets are also vital for developing medical simulators that can be used to train clinicians [5], or train medical robots to perform tasks in realistic anatomical environments [6].

However, creating such datasets, especially for anatomically complex and variable organs like the colon, presents significant c hallenges. S egmenting the target organ from representative medical images is typically labor-intensive, time-consuming, and requires a high level of clinical expertise [2].

Computed Tomography colonography (CTC) is currently the most effective imaging technique for visualizing the 3D anatomy of the colon. This non-invasive method detects and monitors colorectal abnormalities by inflating the colon with a gas, *e.g.,* CO2, and acquiring CTC images [7]. Segmentation from CTC is challenging due to the flexible a natomy o f t he c olon, t he p resence o f o ther gas-filled structures, and the colon variability in size, shape, and p osition. Collapsed segments and residual fluids a dd f urther c omplexity (Figure 2).

We introduce a dataset of 435 human colons, segmented from CTC obtained from the publicly available The Cancer Imaging Archive (TCIA) [8]. Each scan includes a segmentation mask of the whole colon clinically validated by an expert abdominal radiologist, and a segmentation mask of only gas-filled p arts o f the colon.

To our knowledge, this is the first o pen-access d ataset o f s egmented colon anatomies derived from CTC. It provides a valuable resource for population-scale radiologic studies, AI-based image analysis, and the development of digital organ models and simulators, supporting both research and medical device development.

Methods

We used the publicly accessible CTC dataset from TCIA [8]. The dataset includes 825 outpatients aged 50 and older, scheduled for colonoscopy screening with no procedures in the past five y ears. I t c onsists o f 3 ,451 C T s cans w ith a spatial resolution of around 0.8 mm. The dataset includes scans acquired from seven different hospitals. We excluded 1737 scans that exhibited dimensional

inconsistencies, including (1) fewer than 350 or more than 700 axial slices, (2) axial slices smaller than 512×512 pixels, or (3) a disrupted format (Figure 1, *Pre-processing*).

FIGURE 1 GOES HERE

Dataset Annotation

To accelerate colon annotation, we developed a hybrid semi-automatic segmentation pipeline that requires minimal user input. Since the colon in CTC primarily contains gas but may also include fluid, we adopted a two-phase segmentation approach. First, we segmented the gas-filled regions using rules-based traditional methods. Then, we applied interactive machine learning to segment the fluid-filled areas. Finally, the results from both phases were combined to produce the complete colon segmentation.

FIGURE 2 GOES HERE

Semi-automatic annotation of the gas-filled colon segments

Filled with gas, *e.g.,* CO2, the colon appears darker in CTC compared to most of the surrounding organs (Figure 2). To quickly generate a high-quality colon segmentation, we first applied traditional segmentation methods based on simple rules. The images were converted to binary format using a threshold of -800 HU. Fast annotation of gas-filled colon regions was achieved by applying a 3D 26-neighbor region growing algorithm. The seed was placed by automatically extracting a region along the left-right midline, spanning ±50 pixels around the anterior-posterior midpoint, and selecting slices from index 50 to 250 along the inferior-superior axis. The first gas-filled pixel encountered when scanning upward was chosen as the seed, ensuring placement in the rectum in most cases. This selection is based on three key observations: (1) patients are well-centered along the left-right axis but vary along the anterior-posterior axis, (2) the first gas-filled region from an inferior-to-superior scan is likely the colon, and (3) some upper-body scans may include the proximal legs. Cases where automatic seed placement failed were excluded, as this phase aimed for fast colon annotation rather than a universally applicable method (Figure 1, *Segmentation*). To account for potential colon collapse or connections to other organs like the small bowel, segmentations with volumes exceeding 27 cm^3 or below 3.5 cm^3 were discarded. These values were set heuristically, considering the average volume of an inflated colon.

The final colon annotations were firstly validated by a segmentation technical expert, and incorrect ones were removed (*e.g.,* with collapsed segments or incorrectly connected to other organs) (See Figure 1, *Validation* and Figure 4). Initial validation from a technical expert in segmentation - *i.e.,* biomedical engineering with 3 years experience in segmenting organs, with particular focus on the colon - was followed by clinical validation from an expert abdominal

radiologist on the final colon segmentation, as described in Section .

Interactive Machine Learning annotation of fluid-pockets

In CTC, the colon contains fecal residues along with gas, which appear as fluid pockets in the images. This fluid varies in color, covering the full range of CTC pixel intensities, making traditional segmentation methods ineffective (Figure 3). To address this issue, we used an interactive machine learning approach with *RootPainter* [9].

The dataset was prepared by generating enlarged colon masks on the raw images from a subset of the dataset (specifically, the nnU-Net training set) to simplify the segmentation task for the network. This preprocessing step helped focus the attention of the model by excluding irrelevant regions, thereby making the annotation and learning of fluid-filled areas more efficient. To create these coarse colon masks, we used the only open-source tool available for colon segmentation, *TotalSegmentator* [10]. However, since *TotalSegmentator* often produces coarse, low-resolution segmentations, that may exclude parts of the colon, the masks were enlarged with 35 voxels dilation to ensure better coverage (See *fluid annotation example* images in Figure 3). While we could have used dilated versions of our gas-filled colon segmentations to create the masks, we found that *TotalSegmentator* was more effective at capturing the fluid-filled regions. Achieving similar coverage with our own masks would have required significant enlargement, potentially reducing precision.

FIGURE 3 GOES HERE

To create a dataset for training a fluid segmentation model in *RootPainter*, we randomly selected seven axial slices per scan from regions containing the previously segmented gas-filled colon. The created dataset ensures a diverse patient representation and sufficient slices for interactive training. This step resulted in 2,030 slices, which were converted to PNGs with 1000×1000 pixels for detailed annotation and segmentation.

We adopted a 2D model approach instead of a 3D one, as it enables faster feedback, more intuitive user interaction, lower computational cost, and simpler annotation management. These requirements are essential for maintaining an efficient and effective human-in-the-loop workflow. We followed the corrective-annotation protocol described in [9], which involves simultaneous annotation and model training. A segmentation technical expert inspected the model predictions and assigned corrections, which were then incorporated into the training set to refine subsequent predictions. We evaluated the accuracy of the fluid segmentation model with the Dice score, *i.e.,* the difference between the initial predicted segmentation and the corrected segmentation after user annotation is assigned.

The corrective annotation process continued until 1,134 images were evaluated, with annotations assigned to 390 images over 215 minutes. Figure 3 illustrates how the Dice scores improved and fluctuated over time. Annotation stopped

at image 1,134 when *RootPainter*'s colon segmentation was deemed satisfactory upon visual inspection. The Dice also showed diminishing returns, justifying the decision. Performance improved most in the first 80 minutes, reaching a rolling Dice above 0.95 by image 210 (Figure 3). The final trained *RootPainter* model was then used to segment the full dataset, followed by the post-processing described below to refine the results. Components less than 2000 voxels ($\approx 0.002cm^3$) and less than 2 mm from the surface of the colon were removed. Since CTC scanning is typically performed in prone or supine positions, fluid accumulates in the posterior (supine) or anterior (prone) part of the colon due to gravity (see Figure 3). This information was used to refine segmentation by discarding fluid pixels without gas-filled colon pixels within ± 2 axial slices in the axial plane. Additional post-processing steps included hole filling, Gaussian smoothing for surface refinement, and connecting fluid regions to the nearest gas-filled colon by filling empty pixels, in the sagittal plane. The final fluid annotations were again validated by an expert, with additional manual post-processing applied to correct errors, *i.e.,* erase regions that were miss classified as part of the fluid, or fill in missing regions. Figure 4 illustrates an example of segmentation masks both with and without the fluid-filled colon sections.

FIGURE 4 GOES HERE

Data Records

The dataset is hosted in the Open Science Framework (OSF) repository "HQ-Colon: High-Resolution Human Colon Segmentation" [11], as described below.

- The folder *gas-filled-colon-segmentation.zip* contains the 435 segmentation masks of only the gas-filled sections of the colon (*i.e.,* fluid is not included)
- The folder *gas-and-fluid-filled-colon-segmentation.zip* includes the 435 segmentation masks of the entire colon, covering both gas- and fluid-filled regions.
- The file named *meta-data.json* provides the name mapping between our segmentation masks files and the original TCIA images.

All segmentation masks are provided in .mha format, which preserves the metadata from the original CT images in the file headers.

The repository also includes the folder *masks-totalsegmentator.zip*, which contains both the original and dilated (by 35 voxels) segmentation masks generated with *TotalSegmentator* on the same 435 scans [10]. The dilated segmentation masks were used for the interactive machine learning annotation of the fluid pockets.

The segmentation masks of only the fluid-filled regions can be obtained by subtracting the gas-filled colon segmentation labelmap from the gas-and-fluid-filled colon segmentation labelmap. The code for this process is available in the

GitHub repository https://github.com/horizon-europe-2023-ire/colon-segmentation-dataset at *fluid/segmentation/post_processing/subtract_labelmaps.py*.

Naming Convention

Each segmentation follows a standardized naming convention: "colon_XXX" for segmentation masks where "XXX" ranges from 001 to 435. The corresponding original CT images are available from the TCIA CTC repository [8].

The *meta-data.json* file maps each segmentation mask to its corresponding TCIA scan. For each case, it includes the "InstanceUID" of the original image, the subject's sex, scanning position (*i.e.,* prone or supine), and the associated segmentation file name (under the key "nnunet_label_file").

Additionally, the metadata includes three supplementary fields—"subject_id", "number", and "nnunet_image_file"—which may be useful for applications beyond the scope of this paper.

Technical Validation

The technical validation of the segmentation masks was carried out in four stages. First, following the semi-automatic segmentation of the gas-filled section of the colon, all resulting masks were manually reviewed by a segmentation expert, *i.e.,* a biomedical engineer with three years of experience in organ segmentation, with a particular focus on the colon. For each mask, both the overlay of the segmentation on the original image and the corresponding triangular mesh, generated using the marching cubes algorithm, were examined. Any inconsistencies in the masks were manually corrected by removing misclassified gas regions or filling in missing areas. The same technical validation procedure was applied after the semi-automatic generation of the fluid masks (Section), including the correction of misclassified or missing fluid regions. This protocol was also followed after combining the gas-filled and fluid masks. Cases that required excessive post-processing due to large discrepancies between the segmentation and the original image were discarded (n = 51). Finally, the segmentation quality was evaluated by a clinical validation expert, as described in the following section.

Clinical Validation

The quality of the colon segmentations was assessed by an expert abdominal radiologist, *i.e.,* a radiologist with six years of experience in clinical assessment of CTC. Specifically, the clinician was presented with 50 subjects, randomly selected from the full set of 435 segmented CT scans. For each subject, 3D colon reconstructions from both the supine and prone positions were provided. This setup mirrors the conditions clinicians typically encounter in daily practice when reviewing CTC for diagnostic purposes.

The radiologist assessed each colon by indicating their level of agreement with the following five statements, using a 5-point Likert scale: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree:

1. The reconstructed anatomy accurately represents the patient's actual colon anatomy in the CT scans.
2. The reconstruction is sufficient to allow me to assess the level of distension across all segments of the colon.
3. The reconstruction is sufficient to allow me to assess the elongation of all segments of the colon.
4. The reconstruction of haustral folds is comparable in quality to that produced by the software I use in my daily clinical practice (IntelliSense).
5. The overall colon reconstruction is comparable in quality to that produced by the software I use in my daily clinical practice (IntelliSense).

The results show that all subjects received a 'Strongly agree' rating from the expert for the first three questions. For questions 4 and 5, 'Strongly agree' was given for all subjects except in four cases, where the expert responded with 'Disagree'. These findings support the conclusion that the segmentations have been clinically validated by an expert.

Data Availability

The dataset is hosted in the Open Science Framework (OSF) repository "HQ-Colon: High-Resolution Human Colon Segmentation" (https://doi.org/10.17605/OSF.IO/8TKPM).

Code Availability

The code for the semi-automatic segmentation of the colon can be found in the following GitHub repository https://github.com/horizon-europe-2023-ire/colon-segmentation-dataset.
The *RootPainter* project for segmenting the colon fluid is available in the folder "root_painter_colon_fluid_project.zip" of the OSF repository [11].

References

[1] Alabduljabbar, A., Khan, S.U., Alsuhaibani, A., Almarshad, F., Altherwy, Y.N.: Medical imaging datasets, preparation, and availability for artificial intelligence in medical imaging. Journal of Alzheimer's Disease Reports **8**(1), 1471–1483 (2024)

[2] Liu, X., Qu, L., Xie, Z., Zhao, J., Shi, Y., Song, Z.: Towards more precise automatic analysis: a systematic review of deep learning-based multi-organ segmentation. BioMedical Engineering OnLine **23**(1), 52 (2024)

[3] Starck, S., Sideri-Lampretsa, V., Ritter, J.J., Zimmer, V.A., Braren, R., Mueller, T.T., Rueckert, D.: Using uk biobank data to establish population-specific atlases from whole body mri. Communications Medicine **4**(1), 237 (2024)

[4] Tang, C., Yi, W., Occhipinti, E., Dai, Y., Gao, S., Occhipinti, L.G.: A roadmap for the development of human body digital twins. Nature Reviews Electrical Engineering **1**(3), 199–207 (2024)

[5] Finocchiaro, M., Cortegoso Valdivia, P., Hernansanz, A., Marino, N., Amram, D., Casals, A., Menciassi, A., Marlicz, W., Ciuti, G., Koulaouzidis, A.: Training simulators for gastrointestinal endoscopy: current and future perspectives. Cancers **13**(6), 1427 (2021)

[6] Pore, A., Finocchiaro, M., Dall'Alba, D., Hernansanz, A., Ciuti, G., Arezzo, A., Menciassi, A., Casals, A., Fiorini, P.: Colonoscopy navigation using end-to-end deep visuomotor control: A user study. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9582–9588. IEEE (2022)

[7] Mang, T., Graser, A., Schima, W., Maier, A.: Ct colonography: techniques, indications, findings. European journal of radiology **61**(3), 388–399 (2007)

[8] Smith, K., Clark, K., Bennett, W., Nolan, T., Kirby, J., Wolfsberger, M., Moulton, J., Vendt, B., Freymann, J.: Data from ct colonography (2015). https://doi.org/10.7937/K9/TCIA.2015.NWTESAY1

[9] Smith, A.G., Han, E., Petersen, J., Olsen, N.A.F., Giese, C., Athmann, M., Dresbøll, D.B., Thorup-Kristensen, K.: Rootpainter: deep learning segmentation of biological images with corrective annotation. New Phytologist **236**(2), 774–791 (2022)

[10] Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023)

[11] Finocchiaro, M., Stern, R., Ganz, M.: High-resolution human colon segmentation dataset (2025), https://doi.org/10.17605/OSF.IO/8TKPM, Open Science Framework

Author Contributions

Martina Finocchiaro designed and implemented the segmentation methods with contributions from Ronja Stern and Abraham George Smith; performed post-

processing, quality checks and manual corrections of the segmentatios; designed the clinical validation protocol with input from Kristoffer Cold and Lars Konge; designed and developed the interface for clinical validation; analyzed the results and wrote the manuscript. Martina Finocchiaro and Ronja Stern created the data and code repositories. Rikke Vilhelmsborg clinically validated the segmentations. Kenny Erleben advised on the technical content of the research project. Melanie Ganz supervised and coordinated the research project. All authors reviewed and edited the manuscript.

Competing Interests

The author(s) declare no competing interests.

Funding

Funded by the European Union, grant number 101135082. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

FIGURE 5 GOES HERE

Figure legends

1. *File:* Image1.JPG
   Legend: Dataflow for the creation of the colon segmentation Dataset. The grey boxes on the right explain the reasons for scan exclusions at each step, along with the number ("n") of scans discarded for each reason. Age and gender distribution (left) refers to the final colon segmentation dataset.

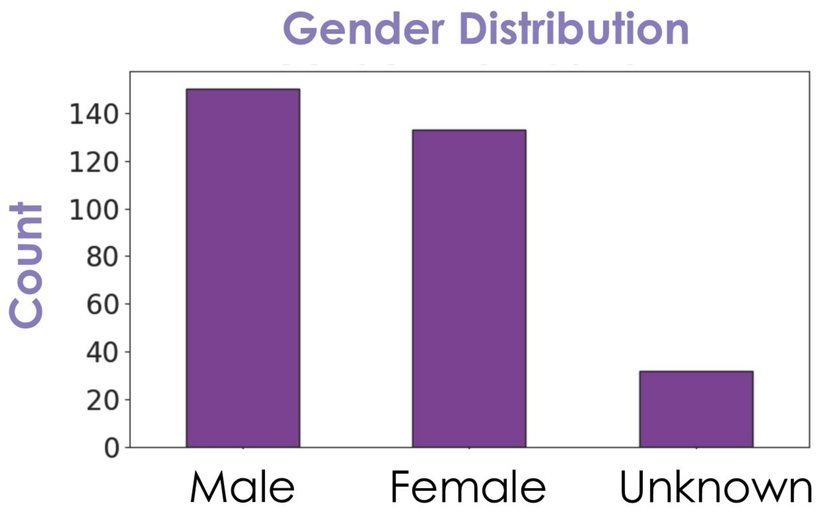
2. *File:* Image2.JPG
   Legend: Example of axial, sagittal, and coronal CT colonography slices (top) with corresponding gas- and fluid-filled colon annotations (bottom). On the right, 3D reconstructions show the colon alone (bottom) - with and without the fluid - and with the small bowel (top). The segmentation task involves 1) annotating the gas-filled regions of the colon and differentiating them from other gas-filled tubular structures, such as the small bowel and lungs; and 2) annotating the intraluminal fluid, that it is challenging, as its intensity can resemble that of bone or muscle. Failure to segment the fluid may result in the loss of important anatomical details, as demonstrated in the 3D models with and without fluid inclusion.

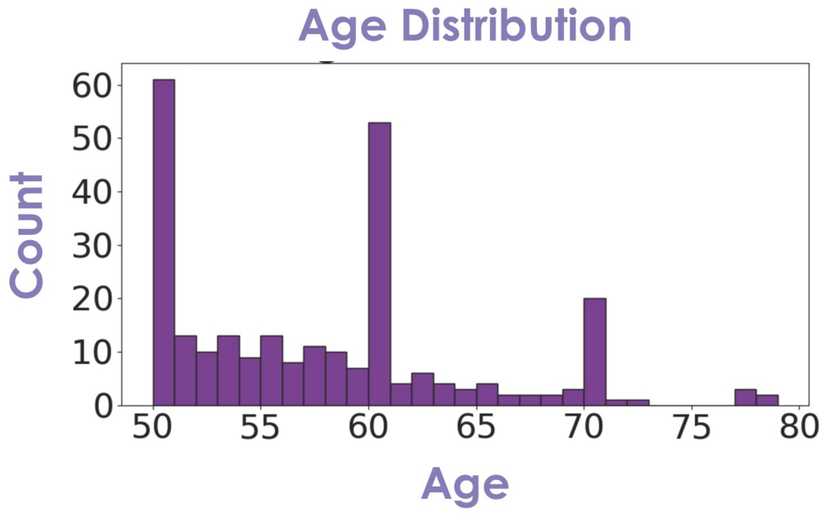
3. *File:* Image3.JPG
   Legend: Examples of fluid visualization on axial slices (top): (a) supine

and (b) prone patient. Images c and d show *RootPainter* fluid segmentation. Dice as a function of the number of annotated images and annotation time (bottom). Observe that after 75 minutes the DICE score of *Rootpainter* flattens out at near 1.

4. *File:* Image4.JPG
   Legend: Example of one colon included in the dataset: 1) original Computer Tomography Colonography, 2) gas-filled colon segmentation mask, 3) full colon segmentation mask, and 4) 3D reconstructed model of both gas-filled and fluid sections. Observe how the presence of fluid influences the segmentation mask, and the corresponding reconstructed 3D model. In its absence, the colon appears flattened on one side due to the effect of gravity pulling fluid downward.

5. *File:* Image5.JPG
   No legend





TCIA Dataset
3451 Scans 825 Subjects

1714 Scans/ 824 Subjects

1708 Scans/ 824 Subjects

Colon segmentation Dataset
435 Scans
315 Subjects

Pre-processing

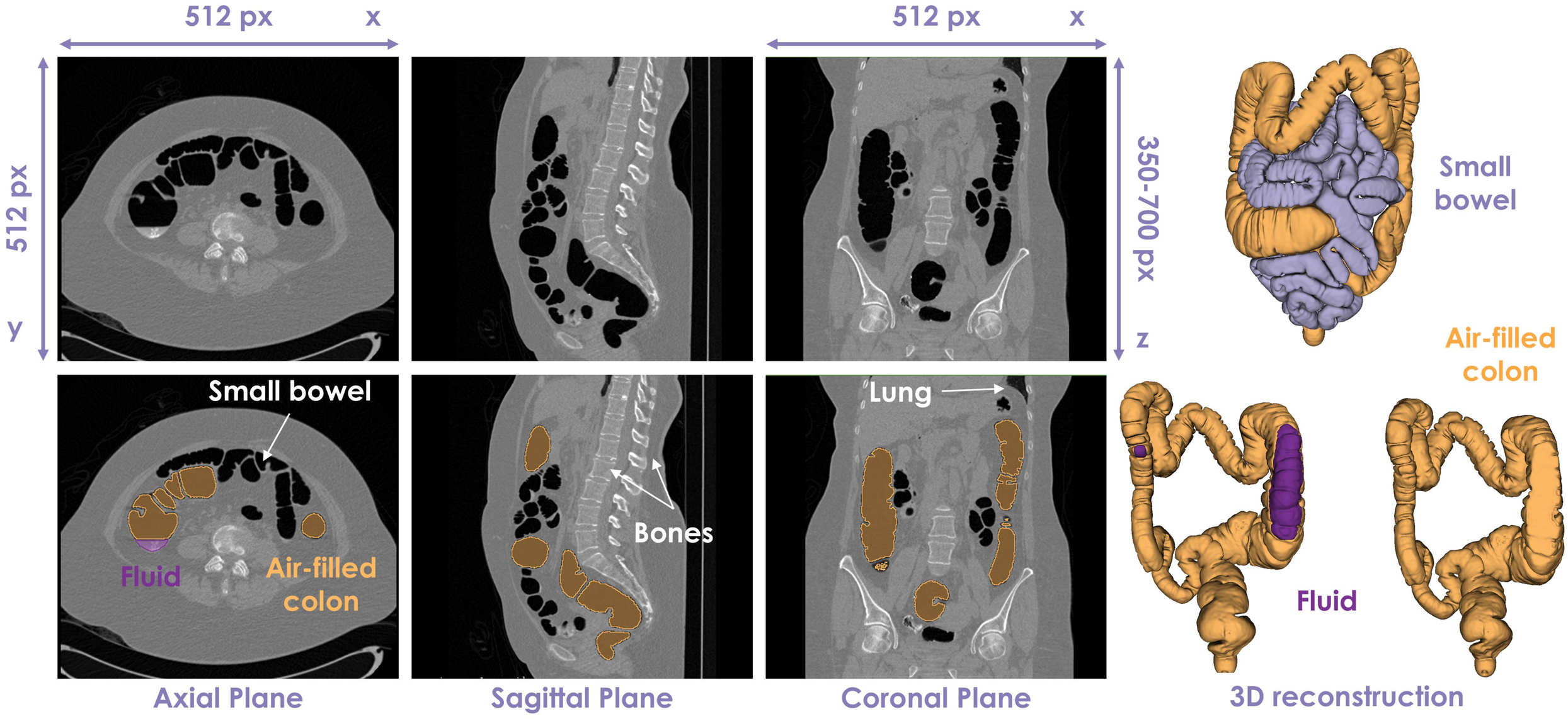
- N. axial slices <350 or >700 (n=1714)
- disrupted format (n=23)

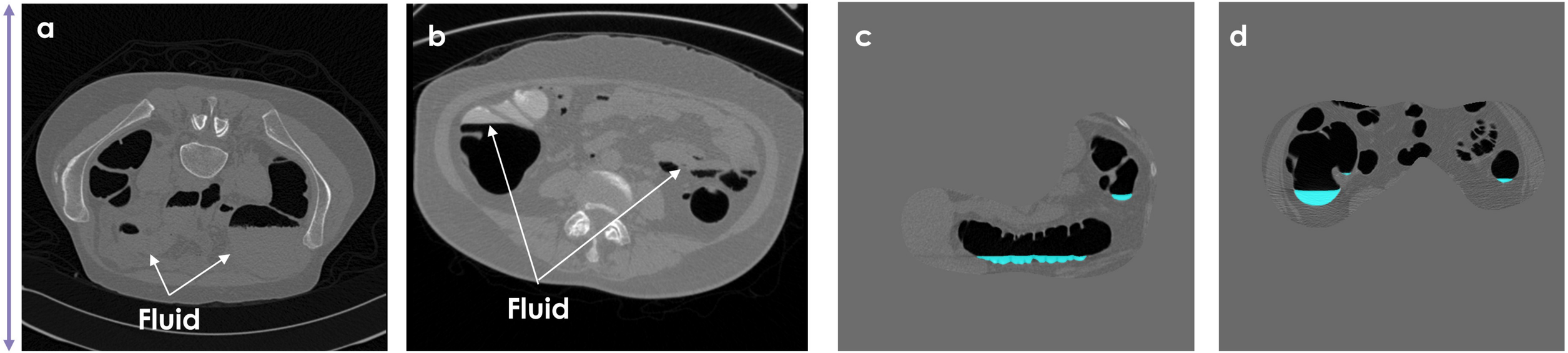
Segmentation

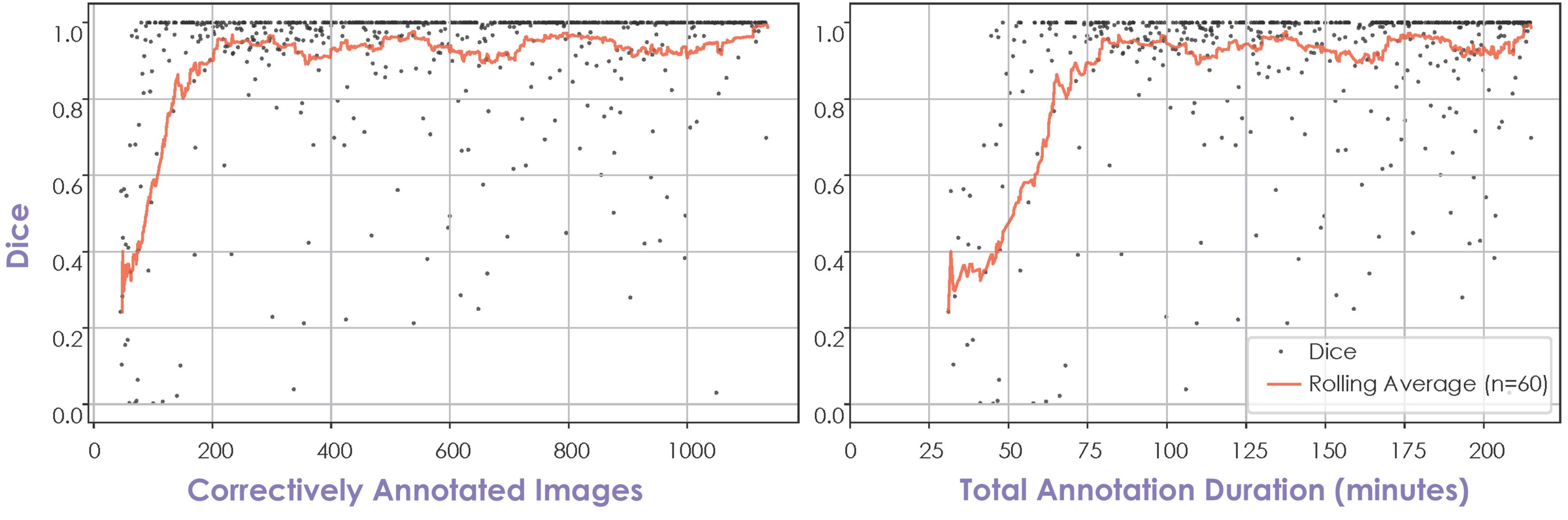
- Automatic region growing seed placement failed (n=6)

Validation

- Segmentation volume <3.5cm or> 27cm (n=1212)
- Expert validation failed (n=51)


512 px
x
512 px
y
512 px
x
350-700 px
z
Small bowel
Air-filled colon
Small bowel
Fluid
Air-filled colon
Bones
Lung
Fluid
Axial Plane
Sagittal Plane
Coronal Plane
3D reconstruction


a
Fluid
b
Fluid
c
d
Supine
Prone
Fluid annotation examples


Dice
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
Correctively Annotated Images
0
25
50
75
100
125
150
175
200
Total Annotation Duration (minutes)
Dice
Rolling Average (n=60)

Original

Gas-filled colon segmentation

Fluid missing

Full colon segmentation

Fluid included

3D Renconstructed model

Fluid

Front

Lateral

Back

Funded by
the European Union