

1 Development of a cloud framework for training and deployment of deep learning models in
2 Radiology: automatic segmentation of the human spine from CT-scans as a case-study
3 Rui Santos^{1*}, Nicholas Büniger^{1*}, Benedikt Herzog^{1,.,} and Sebastiano Caprara^{1#}

4
5 ¹Digital Medicine Unit, Balgrist University Hospital Zürich, University of Zürich, Forchstrasse 340,
6 8008, Zurich, CH, Switzerland.

7 *Equally contributing authors

8 #Corresponding author: sebastiano.caprara@balgrist.ch

9

10

11 **Abstract**

12 Advancements in artificial intelligence (AI) and the digitalization of healthcare are revolutionizing
13 clinical practices, with the deployment of AI models playing a crucial role in enhancing
14 diagnostic accuracy and treatment outcomes. Our current study aims at bridging image data
15 collected in a clinical setting, with deployment of deep learning algorithms for the segmentation
16 of the human spine. The developed pipeline takes a decentralized approach, where selected
17 clinical images are sent to a trusted research environment, part of private tenant in a cloud
18 service provider. As a use-case scenario, we used the TotalSegmentator CT-scan dataset,
19 along with its annotated ground-truth spine data, to train a ResSegNet model native to the
20 MONAI-Label framework. Training and validation were conducted using high performance
21 GPUs available on demand in the Trusted Research Environment. Segmentation model
22 performance benchmarking involved metrics such as dice score, intersection over union,
23 accuracy, precision, sensitivity, specificity, bounding F1 score, Cohen's kappa, area under the
24 curve, and Hausdorff distance. To further assess model robustness, we also trained a state-of-
25 the-art nnU-Net model using the same dataset and compared both models with a pre-trained
26 spine segmentation model available within MONAI-Label. The ResSegNet model, deployable
27 via MONAI-Label, demonstrated performance comparable to the state-of-the-art nnU-Net
28 framework, with both models showing strong results across multiple segmentation metrics. This
29 study successfully trained, evaluated and deployed a decentralized deep learning model for CT-
30 scan spine segmentation in a cloud environment. This new model was validated against state-
31 of-the-art alternatives. This comprehensive comparison highlights the value of the MONAI-Label
32 as an effective tool for label generation, model training, and deployment, further highlighting its
33 user-friendly nature and ease of deployment in clinical and research settings. Further we also
34 demonstrate that such tools can be deployed in private and safe decentralized cloud
35 environments for clinical use.

36

37 Key words: U-Net, ResSegNet, nnU-Net, deep-learning, framework, MONAI, cloud computing

38

39 Abbreviations:

40 DL: deep learning; AI: artificial intelligence; AUC: area under the curve

41

42 **Author Summary:**

43 In the rapidly evolving field of medical imaging, the integration of artificial intelligence (AI) and
44 cloud computing is becoming increasingly critical for advancing diagnostic and treatment
45 capabilities. To address the growing demand for flexible digital frameworks in the clinical
46 environment supporting the deployment of data-driven applications, we have developed and
47 deployed a cloud-based Trusted Research Environment designed specifically for training,
48 validation, and deployment of deep learning models focused on semantic segmentation in
49 musculoskeletal radiology. This environment facilitates the efficient handling of large datasets
50 and the accessibility of algorithmic output for the physicians, optimizing the interface between
51 development and clinical translation. The established framework enables significant
52 improvements in the deployment of deep learning tools for image analysis in the clinical setting.
53 In our current use-case, we have utilized this environment to train and evaluate two advanced
54 deep learning models for the segmentation of the human spine from CT scans. By leveraging
55 the computational power and flexibility of the cloud-based infrastructure, we were able to
56 perform rigorous training and comparison of these models, aiming to enhance the accuracy and
57 reliability of spine segmentation in clinical practice. This approach not only streamlines the
58 process of model development but also provides valuable insights into the performance and
59 potential clinical applications of these AI-driven segmentation tools.

60

61 **Introduction**

62 With the inflection point of Artificial Intelligence (AI) algorithms, with its deep learning (DL)
63 models and frameworks (1), medical operations are facing the paradigm shift of both
64 digitalization and DL models deployment. Out of all clinical specialties, Radiology is poised to be
65 at the forefront of this transformation due to its inherent reliance on visual data and the well-
66 established role of image analysis in diagnosis and treatment planning (2,3). As such, there is a
67 strong research focus on developing and deploying AI models within the Radiology field. As an
68 example, in the past 5 years (2019-2023), a total of 11'392 research manuscripts (PubMed
69 search) were related in some form with the thematic of radiology and deep learning (DL).
70 Indeed, medical imaging is a key area for the development of high added-value DL models that
71 streamline and accelerate the process of image segmentation and classification, which is a
72 major time-consuming task for physicians (4).
73 Several AI tools have been developed with the objective of accelerating medical image
74 segmentation. Recently, the development of the TotalSegmentator DL model (5), trained to
75 segment 117 anatomical structures (version 2.0), from CT datasets, demonstrates the immense

76 possibilities that DL-powered medical image segmentation offers. For the development of such
77 a powerful DL segmentation model, the nnU-Net framework was used (6). This framework
78 allows the training of state-of-the art segmentations models, based on the U-Net DL architecture
79 (7), with minimal to no coding needed. Still, these tools are limited in their deployment in daily
80 clinical scenarios. As an easier and user-oriented approach, the open-source framework MONAI
81 (Medical Open Network for Artificial Intelligence) (8) has been developed for the training and
82 easier deployment of automatic medical image segmentation models. At the forefront of MONAI,
83 the sub-framework MONAI-Label (9) connects DL architectures and pretrained models with
84 open-source graphic user interfaces (GUI) such as 3D Slicer and OHIF (Open Health Imaging
85 Foundation Viewer), enabling active learning with direct visualization and correction of
86 segmentation results (10,11).

87 As optimal environment for the deployment of such aforementioned frameworks, the adoption of
88 cloud computing in medical research setting, offers many benefits: it fosters improved
89 collaboration among physicians and researchers, provides a decentralized and scalable
90 computing infrastructure and greater accessibility to tools and services such as giving support for
91 data annotation. Furthermore, cloud-based infrastructure offers simplified management of
92 resources in a cost-effective manner, allowing medical research centers to focus resources on
93 data analysis and clinical relevance of research questions rather than IT maintenance (12,13).

94 In this study, we have developed a cloud computing framework leveraging the open-source
95 pipeline of MONAI-Label for the development and deployment for clinical research teams of an
96 interactive spine segmentation model out of CT data. Moreover, we benchmarked the model
97 output with an in-house trained nnU-Net model, both trained and validated with the same
98 collection of images (TotalSegmentator open-source dataset). We also compare it with a
99 pretrained model for spine segmentation present in MONAI-Label, trained on the VERSE
100 dataset (14). We demonstrate that the MONAI-Label developed model produces similar
101 segmentation results to the model trained with the nnU-Net framework yet providing a far
102 simpler vertical integration in a clinical research setting. This *de novo* trained model also
103 considerably outperforms the pretrained spine segmentation model present in MONAI-Label *ab*
104 *initio*. Moreover, the presented research highlights the use of MONAI-Label as a powerful tool
105 for developing DL models and automatically generating anatomical segmentation of anatomical
106 structures which may support surgical planning, diagnostics and future creation of digital twins
107 of patients (15,16).

108

109 **Materials and Methods**

110 *Architecture of a trusted cloud research environment for deep learning model training and*
111 *deployment*

112 The framework introduced in this work is based on a private cloud environment composed of
113 private endpoints, i.e., a pointer to the location of the data, which allows the hospital network to
114 be extended in a private cloud space (figure 1A). Through an Application Programming Interface
115 (API) Manager, the DICOM (Digital Imaging and Communications in Medicine standard) or NIFTI
116 (Neuroimaging Informatics Technology Initiative) storage locations connect with cloud
117 computational nodes (with adequate hardware accelerators, such as graphic processing units –
118 GPUs), and a centralized management dashboard that allows the definition of strict governance
119 policies, as the ones in place at healthcare institutions, granting secure and private access to
120 specified users. Through the combination of the cloud services mentioned above, the framework
121 composed of 3D Slicer and MONAI Label-based computational nodes was developed. In this way,
122 scalable computation resources are connected with a Graphical User Interface (GUI) thanks to
123 the open-source tool 3D Slicer, as shown on figure 1b.

124 For segmentation model training in the trusted research environment, the TotalSegmentator
125 open-source image dataset was stored in a project-specific NIFTI storage and connected with
126 computational resources embedded in the computational node via the API Manager.

127 For segmentation inference, an additional API allows the access to DICOM images stored in the
128 clinical PACS (Picture Archiving and Communication System), retrieve these images to a DICOM
129 storage location in the cloud which then directly connects to the MONAI Label computational node
130 for automatic segmentation processing. The connection to the GUI results in a user-friendly
131 environment where radiologists and other physicians can visualize the segmentation results,
132 correct them, and potentially fine-tuning the models by sending the corrected results back to
133 MONAI Label central instance to update the training.

134

135 *Resources for deep learning model training*

136 The TotalSegmentator dataset composed of CT images and their ground-truth vertebrae and
137 sacrum segmentations (part of the open-source dataset) was used for model training in either the
138 MONAI-Label or nnU-Net frameworks (5) (<https://zenodo.org/records/10047292>, with an average
139 isotropic pixel size of 1.5 mm). The dataset was divided into training and inference dataset. 964
140 images were randomly chosen to be part of the training dataset, while the remaining 108 were
141 used for validation (the open-source version of the dataset contains more than 1200 images,
142 nonetheless, some cannot be used with the current version, 2.0.1). The dataset was
143 preprocessed in accordance with the requirements of both MONAI-Label and nnU-Net

144 (instructions found at [https://github.com/Project-MONAI/MONAILabel?tab=readme-ov-](https://github.com/Project-MONAI/MONAILabel?tab=readme-ov-file#getting-started-with-monai-label)
145 [file#getting-started-with-monai-label](https://github.com/Project-MONAI/MONAILabel?tab=readme-ov-file#getting-started-with-monai-label) and <https://github.com/MIC-DKFZ/nnUNet>). The dataset
146 was uploaded into a storage location in the developed cloud-based Trusted Research
147 Environment. The default segmentation task in MONAI-Label , based on a ResSegNet
148 architecture (17), was used for spine segmentation model training. The training was performed in
149 4 A100 GPU computational nodes, for 1000 epochs, 241 iterations each, taking *circa* 25 hours of
150 compute time. Since for both training and results analysis a graphic user interface is necessary,
151 the open-source 3D Slicer tool was used in a container (<https://github.com/pieper/SlicerDockers>,
152 v5.0.3). No validation split was used in this training as model benchmarking would be done with
153 an already assigned dataset (remaining 108 images available). Default segmentation pipeline
154 scripts were used with no modifications (segmentation.py, part of the radiology app from the
155 MONAI-Label framework with a tutorial found here: [https://github.com/Project-](https://github.com/Project-MONAI/tutorials/tree/main/monailabel)
156 [MONAI/tutorials/tree/main/monailabel](https://github.com/Project-MONAI/tutorials/tree/main/monailabel)). The nnU-Net model training was performed on one A100
157 compute node for 1000 epochs, 250 iterations each, over the course of 15 hours of training. No
158 validation split was used (fold 0), as model benchmarking would be done with an already assigned
159 dataset (same 108 datasets used to validate the MONAI-Label model). Default nnU-Net pipeline
160 scripts were used with no modifications, following the developers tutorial.

161

162 *Deep learning model benchmarking*

163 108 CT scans of the TotalSegmentator dataset (not part of the dataset for DL model training)
164 were used for model validation. We calculated different segmentation metrics to benchmark both
165 deep learning models' segmentation results such as dice score, intersection over union, accuracy,
166 precision, sensitivity, specificity, bounding F1 score, recall, area under the curve (AUC) and the
167 Hausdorff distance between segmentation predictions and their ground truth (18). Each metric
168 was calculated with custom scripts (available upon request). Calculation of distances between
169 ground truth and predicted segmentations was done in the open-source software Meshlab
170 (<https://www.meshlab.net/>) v2022.02 with built-in functions. P-values were calculated based on a
171 non-paired t-student function.

172

173 **Results**

174 *Benefits of a decentralized cloud infrastructure for DL model training and deployment*

175 Deploying a decentralized trusted research environment on the cloud for training and deployment
176 of deep learning models present two major benefits: on one hand the scalability and flexibility to

177 compose a tailored DL model for a specific research or clinical question, on the other hand, the
178 possibility to deploy the developed models in clinical workflows using the same architecture.
179 Depending on the size and architecture of the model, the compute resources of the computational
180 nodes can be quickly scaled-up or scaled-down, providing the creation of an optimized setup for
181 each use-case requiring a training, validation and/or deployment of a research or clinical DL
182 model. Further, this decentralized approach makes it far simpler to securely and privately connect
183 different medical devices (CT scanners, MRI machines, ultrasound machines, etc.) as endpoints
184 to the developed trusted research environment and get outputs in real-time.
185 Within the MONAI Label framework, multiple models can easily be retrieved and fine-tuned by
186 labeling additional data via a user-friendly GUI, in this way a local version can be trained. The
187 versatility of this process enables various healthcare operators, e.g. radiologists, to retrain models
188 on top of open-source available models (transfer learning). Furthermore, they can easily test
189 models that have been previously created in the framework and validate them thanks to
190 commonly available software for 3D visualization. This way, the interaction between model's
191 development and radiologists is considerably improved and physicians are a central component
192 in the development of new ML models.

193
194 *Training a ResSegNet based CT-scan spine segmentation model within the MONAI-Label*
195 *framework: spineCT-ResSegNet model*

196 Out of 1072 images present in the TotalSegmentator, 964 images (90% of the dataset), were
197 randomly chosen to compose the training dataset, while 108 (remaining dataset) served as the
198 validation dataset. Default 3D ResSegNet parameters of the segmentation (ResSegNet) task of
199 MONAI-Label radiology app, were used. Qualitative analysis of the segmentation results (figure
200 2A) demonstrates the capacity of robust prediction of the entire human spine and sacrum (25
201 segmented classes). Vertebral bodies identification is correctly done, with their shape being in
202 accordance with anatomical expectancy. Quantitative segmentation quality metrics over all the
203 25 classes reveal a robust model for spine segmentation from CT-scans with an average dice
204 score of 0.856 ± 0.066 , an average intersection over union score of 0.801 ± 0.074 , average
205 accuracy of 1.000 ± 0.000 , a mean precision of 0.862 ± 0.066 , a mean sensitivity (or recall) of
206 0.861 ± 0.070 , a mean specificity of 1.000 ± 0.000 , an average bounding F1 score of 0.837 ± 0.057 ,
207 a mean area under the curve (AUC) of 0.932 ± 0.034 and a Kappa of 0.853 ± 0.068 (figure 2B, all
208 numbers rounded to 3 decimal places). Detailed results, with 95% confidence intervals, are found
209 in supplementary figure 1.

210

211 *Training a nnU-Net based CT-scan spine segmentation model: spineCT-nnUNet model*

212 Serving as a benchmark for DL segmentation model comparison, we trained the nnU-Net
213 framework on the same dataset [OBJ:OBJ], with the 3D-UNet full resolution model (1.5 mm), using the
214 default configuration and trained the model over 1000 epochs. Validation was performed with the
215 same 108 CT scans as for the MONAI-Label ResSegNet model, native to MONAI-Label.
216 Segmentation output is qualitatively robust with anatomical structures, i.e. vertebrae, are localized
217 to their correct position and have a correct tridimensional conformational (figure 3A). [OBJ:OBJ] between
218 the segmentation prediction and their ground-truth, reveal a robust prediction model, with an
219 average dice score over the 25 classes (anatomical structures) of 0.914 ± 0.036 , an average
220 intersection over union of 0.879 ± 0.045 , an accuracy of 1.000 ± 0.000 , a precision of 0.911 ± 0.046 ,
221 a sensitivity (or recall) of 0.910 ± 0.050 , a specificity of 1.000 ± 0.000 , a boundary F1 score of
222 0.908 ± 0.029 , an AUC score of 0.961 ± 0.018 and a Kappa score of 0.904 ± 0.049 (figure 3B, 5 all
223 numbers rounded to 3 decimal places). Detailed results, with 95% confidence intervals, are found
224 in supplementary figure 2.

225

226 *Comparison of spineCT-ResSegNet, spineCT-nnUNet and pretrained spine segmentation*
227 *models*

228 To further compare the segmentation predictions between the ResSegNet model and the nnU-
229 Net model to the respective ground truth, we overlaid selected examples as exemplified in figure
230 4A (spineCT-ResSegNet) and 4C (spineCT-nnUNet). Hausdorff distances between each
231 segmentation model predictions and their ground truths were calculated (spineCT-ResSegNet
232 and spineCT-nnUNet in figure 4B and figure 4D respectively). The spineCT-ResSegNet model
233 presents an average Hausdorff distance of 2.604 ± 0.450 mm while the spineCT-nnUNet model
234 predictions present a Hausdorff distance of 2.214 ± 0.398 . While the nnU-Net based model
235 outperforms the ResSegNet model in regard to the quality metric of Hausdorff distance, the same
236 is not observed for the remaining metrics calculated. Thus, we find sufficient evidence to say both
237 models produce on-par segmentation results (supplementary figure 5A, all p-valued rounded to 5
238 decimal plates).

239 We also benchmarked the pretrained vertebrae segmentation model which is part of the MONAI-
240 Label core. The model spine segmentation predictions average a dice score over the 24 classes
241 (as this model only predicts from C1 to L5) of 0.601 ± 0.153 , an average intersection over union of
242 0.533 ± 0.154 , an accuracy of 1.000 ± 0.000 , a precision of 0.697 ± 0.150 , a sensitivity (or recall) of
243 0.565 ± 0.157 , a specificity of 1.000 ± 0.000 , a boundary F1 score of 0.588 ± 0.144 , an AUC score
244 of 0.785 ± 0.080 and a Kappa score of 0.597 ± 0.151 (supplementary figure 3, all numbers rounded

245 to 3 decimal places). The output segmentations present a Hausdorff distance of 3.129 ± 0.322 mm
246 on average. Thus, we can demonstrate that the *de novo* trained model spineCT-ResSegNet
247 outperforms the pretrained model in MONAI-Label in most quality metrics calculated
248 (supplementary figure 5B, all p-values rounded to 5 decimal plates). Furthermore, it also trained
249 in one more class, the sacrum.

250

251 **Discussion**

252 A decentralized trusted research environment as a core of novel digital ecosystems in medical
253 research has the potential to simplify communication between physicians and researchers, enable
254 effective testing and validation of novel tools, provide secure and private data exchange channels
255 - and ultimately reduce costs by sharing computational resources, with non-complicated
256 scalability. Further benefits are the ability to access common datasets, shared code, and
257 validation methods at any time reduces errors and improves usability - and thus accelerates the
258 clinical translation of digital technologies. Altogether, such platforms facilitate testing and
259 validation of digital and DL pipelines, making data-driven medical research more reproducible and
260 better suited for the deployment in clinical settings. With these principles in mind, we developed
261 a radiology-specific cloud infrastructure, based on a combination of DICOM storage services and
262 MONAI-Label-compatible computational nodes for the training and deployment of deep learning
263 segmentation models.

264 As a case-study, we trained and deployed a ResSegNet model for human spine segmentation
265 out the open-source dataset of CT-scans from TotalSegmentator. This model was trained as part
266 of the MONAI-Label (8,19) framework. As a benchmark, we used the nnU-Net framework (6) and
267 also trained a segmentation model with the same training and inference dataset. Both models
268 yielded similar quality metric results with only average Hausdorff distance being noticeable better
269 for the nnU-Net model, than for the ResSegNet model (figure 4 and supplementary figure 5).
270 Nonetheless, given the native deployment of the ResSegNet model via MONAI-Label, we can
271 directly deploy it via a graphic user interface such as 3D Slicer or OHIF (10,11), which is not
272 possible with the nnU-Net, at least not in a straightforward way. This in-place visualization
273 provides direct access to the segmentation results to the health professionals who can use it to
274 make informed decisions, e.g. for surgical planning.

275 We also compared our trained model, spineCT-ResSegNet, with the pre-trained and pre-built
276 spine segmentation model which MONAI-Label already offers (14). Our newly trained model
277 greatly outperforms the prebuilt model in all metrics, providing substantially more accurate
278 segmentations (supplemental figure 4 and 5).

279 While our developed models underperformed compared with the dice scores reported for the
280 TotalSegmentator model (average 0.95 for vertebrae) (5), this nnU-Net based model was trained
281 for 8000 epochs compared to ours which was trained for only 1000 epochs. This meant a far lower
282 training time for our model, fewer computational resources were needed, while still performing
283 adequately for the task at hand.

284 Naturally, given the dice coefficient of around 0.85 and a hausdorff coefficient of approximately
285 2.6 mm, the ResSegNet model incurs in mispredictions and proper segmentation correction may
286 be necessary. Nonetheless, the iterative nature and easy of use of MONAI-Label and a chosen
287 GUI such as 3d Slicer, makes any necessary corrections easy to perform. Further, this corrections
288 may be included as part of a subsequent finetuned model, as part of the labelling correction
289 process, which will then, increase the predictive power of the model.

290

291 **Conclusion**

292 The present research demonstrates the potential of cloud-based deep learning for training and
293 deployment of medical image segmentation models, in a trusted research environment. By
294 leveraging the efficiency and scalability of cloud computing, we have created a framework that
295 empowers medical professionals, such as radiologists, with faster, more accessible segmentation
296 tools, based on the MONAI-Label open-source framework. This paves the way for improved
297 diagnostics, treatment planning, and ultimately, better patient outcomes.

298

299 **Author Contribution**

300 NB developed and deployed the trusted cloud environment. RS worked on the development and
301 testing of the deep learning segmentation models. BH assisted NB in the infrastructure setting of
302 the trusted cloud environment. NB, RS and SC wrote the manuscript and designed the figures.
303 SC supervised the design of the infrastructure and pipeline and contributed to the design of the
304 research and interpretation of the results. MF contributed to the final version of the manuscript.

305

306 **Acknowledgements**

307 The authors acknowledge the Digital Society Initiative, University of Zurich. This work is part of
308 the DSI-Infrastructure / Lab-Program.

309

310 **References**

- 311 1. Sevilla J, Heim L, Ho A, Besiroglu T, Hobbhahn M, Villalobos P. Compute Trends Across Three Eras
312 of Machine Learning. Proceedings of the International Joint Conference on Neural Networks.
313 Institute of Electrical and Electronics Engineers Inc.; 2022. doi:
314 10.1109/IJCNN55064.2022.9891914.
- 315 2. Saba L, Biswas M, Kuppili V, et al. The present and future of deep learning in radiology. Eur J
316 Radiol. Elsevier Ireland Ltd; 2019. p. 14–24. doi: 10.1016/j.ejrad.2019.02.038.

- 317 3. Chea P, Mandell JC. Current applications and future directions of deep learning in
318 musculoskeletal radiology. *Skeletal Radiol.* Springer; 2020. p. 183–197. doi: 10.1007/s00256-019-
319 03284-z.
- 320 4. Langlotz CP. The Future of AI and Informatics in Radiology: 10 Predictions. *Radiology.* Radiological
321 Society of North America Inc.; 2023. doi: 10.1148/radiol.231114.
- 322 5. Wasserthal J, Breit HC, Meyer MT, et al. TotalSegmentator: Robust Segmentation of 104
323 Anatomic Structures in CT Images. *Radiol Artif Intell.* Radiological Society of North America Inc.;
324 2023;5(5). doi: 10.1148/RYAI.230024/ASSET/IMAGES/LARGE/RYAI.230024.VA.JPEG.
- 325 6. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for
326 deep learning-based biomedical image segmentation. *Nature Methods* 2020 18:2. Nature
327 Publishing Group; 2020;18(2):203–211. doi: 10.1038/s41592-020-01008-z.
- 328 7. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image
329 segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial
330 Intelligence and Lecture Notes in Bioinformatics).* Springer Verlag; 2015. p. 234–241. doi:
331 10.1007/978-3-319-24574-4_28.
- 332 8. Jorge Cardoso M, Li W, Brown R, et al. MONAI: An open-source framework for deep learning in
333 healthcare. 2022;
- 334 9. Diaz-Pinto A, Alle S, Nath V, et al. MONAI Label: A framework for AI-assisted Interactive Labeling
335 of 3D Medical Images. 2022; doi: 10.1016/j.media.2024.103207.
- 336 10. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the
337 Quantitative Imaging Network. *Magn Reson Imaging.* 2012;30(9):1323–1341. doi:
338 10.1016/j.mri.2012.05.001.
- 339 11. Ziegler E, Urban T, Brown D, et al. Open Health Imaging Foundation Viewer: An Extensible Open-
340 Source Framework for Building Web-Based Imaging Applications to Support Cancer Research. *JCO
341 Clin Cancer Inform.* 2020. <https://doi.org/10.1016/j.media.2024.103207>.
- 342 12. Griebel L, Prokosch HU, Köpcke F, et al. A scoping review of cloud computing in healthcare. *BMC
343 Med Inform Decis Mak.* BioMed Central Ltd; 2015. doi: 10.1186/s12911-015-0145-7.
- 344 13. Navale V, Bourne PE. Cloud computing applications for biomedical science: A perspective. *PLoS
345 Comput Biol.* Public Library of Science; 2018;14(6). doi: 10.1371/journal.pcbi.1006144.
- 346 14. Sekuboyina A, Hussein ME, Bayat A, et al. VERSE: A Vertebrae labelling and segmentation
347 benchmark for multi-detector CT images. *Med Image Anal.* Elsevier B.V.; 2021. doi:
348 10.1016/j.media.2021.102166.
- 349 15. Caprara S, Carrillo F, Snedeker JG, Farshad M, Senteler M. Automated Pipeline to Generate
350 Anatomically Accurate Patient-Specific Biomechanical Models of Healthy and Pathological FSUs.
351 *Front Bioeng Biotechnol.* Frontiers Media S.A.; 2021;9. doi: 10.3389/fbioe.2021.636953.

- 352 16. Iqbal JD, Krauthammer M, Biller-Andorno N. The Use and Ethics of Digital Twins in Medicine.
353 Journal of Law, Medicine and Ethics. Cambridge University Press; 2022;50(3):583–596. doi:
354 10.1017/jme.2022.97.
- 355 17. Saxena N, Babu NK, Raman B. Semantic segmentation of multispectral images using res-seg-net
356 model. Proceedings - 14th IEEE International Conference on Semantic Computing, ICSC 2020.
357 Institute of Electrical and Electronics Engineers Inc.; 2020. p. 154–157. doi:
358 10.1109/ICSC.2020.00030.
- 359 18. Müller D, Soto-Rey I, Kramer F. Towards a guideline for evaluation metrics in medical image
360 segmentation. BMC Res Notes. BioMed Central Ltd; 2022. doi: 10.1186/s13104-022-06096-y.
- 361 19. Diaz-Pinto A, Alle S, Nath V, et al. MONAI Label: A framework for AI-assisted Interactive Labeling
362 of 3D Medical Images. 2022; <http://arxiv.org/abs/2203.12362>.
- 363
- 364

365 **Figure Legends**

366

367 **Figure 1: Architecture of the cloud infrastructure behind the trusted research environment**

368 **and deployed framework with 3D Slicer and MONAI-Label. (A)** Overview of the components
369 of the architecture of the trusted research environment: storage and compute. The first one has
370 a DICOM blob storage which is exposed with a private endpoint; while the second one has more
371 components. The compute workspace is exposed as well with a private endpoint, while the other
372 component such as storage account, key vault Jupyter notebook and container registry are
373 connected to the workspace. Both systems have group permissions that can be managed
374 individually. To access the trusted research environment, users can use a Virtual Desktop
375 application **(B)** Using two compute nodes, 3D Slicer and MONAI Label are deployed separately.
376 With an SSH connection users can access an online environment to use the 3D Slicer interface
377 to interact with the model's output that runs on the first compute node (from the left side), which
378 process the DICOM images retrieved by the second compute node on which MONAI Label is
379 running from the DICOM blob storage.

380

381 **Figure 2 – SpineCT-ResSegNet spine segmentation model predictions and segmentation**

382 **metrics. (A)** Example of segmentation results with the spineCT-ResSegNet, a trained MONAI-
383 Label native model, in both 2D and 3D visualizations. Vertebrae are color coded. The entire
384 human spine is represented in the image: cervical spine (C1-C7), thoracic spine (T1-T12),
385 Lumbar spine (L1-L5) and sacrum. **(B)** Validation metrics for spineCT-ResSegNet spine
386 segmentation model, based on the prediction of 108 cases. Color encodes the accuracy of
387 segmentation class predictions. Blue indicates high similarity to the ground truth, while red
388 represents increasing deviation from the actual values.

389

390 **Figure 3 – SpineCT-nnUNet spine segmentation model predictions and segmentation**

391 **metrics. (A)** Example of segmentation results with the nnU-Net framework derived model (3D-
392 full resolution) in both 2D and 3D visualizations. Vertebrae are color coded. The entire human
393 spine is represented in the image: cervical spine (C1-C7), thoracic spine (T1-T12), Lumbar
394 spine (L1-L5) and sacrum. **(B)** Validation metrics for spineCT-nnUNet spine segmentation
395 model. Color encodes the accuracy of segmentation class predictions. Blue indicates high
396 similarity to the ground truth, while red represents increasing deviation from the actual values.

397

398 **Figure 4 – Comparison of spineCT-ResSegNet and spineCT-nnUNet semantic**
399 **segmentation models.** (A) Example of the Hausdorff distance calculation between the spineCT-
400 ResSegNet model (native to MONAI-Label) predictions and ground truth. (B) Average Hausdorff
401 distance calculation for spineCT-ResSegNet predictions versus their ground truth. Predictions
402 with the spineCT-ResSegNet model average a Hausdorff Distance of 2.604 ± 0.450 mm to the
403 ground truth. (C) Example of the Hausdorff distance calculation between the spineCT-nnUNet
404 model predictions and ground truth. (D) Average Hausdorff distance calculation for spineCT-
405 nnUNet predictions versus their ground truth. Predictions with the spineCT-nnUNet model
406 average a Hausdorff Distance of 2.214 ± 0.398 mm to the ground truth. Color encoding in (B) and
407 (D) correlates with the accuracy of segmentation class predictions. Blue indicates high similarity
408 to the ground truth, while red represents increasing deviation from the actual values.

409
410 **Supplemental Figure 1 – Comprehensive metrics evaluation of the spineCT-ResSegNet.**
411 Average metrics for the spineCT-ResSegNet model based on comparison between the
412 predictions of the model on 108 datasets and their ground truth.

413
414 **Supplemental Figure 2 – Comprehensive metrics evaluation of the spineCT-nnUNet.**
415 Average metrics for the spineCT-ResSegNet model based on comparison between the
416 predictions of the model on 108 datasets and their ground truth.

417
418 **Supplemental Figure 3 – Comprehensive metrics evaluation of the open-source vertebrae**
419 **segmentation model of MONAI-Label.** Average metrics for the spineCT-ResSegNet model
420 based on comparison between the predictions of the model on 108 datasets and their ground
421 truth.

422
423 **Supplementary Figure 4 – Hausdorff distance between prediction and ground truth by the**
424 **open-source vertebrae segmentation model of MONAI-Label.** (A) Example of the Hausdorff
425 distance calculation between the spineCT-ResSegNet model (native to MONAI-Label) predictions
426 and ground truth. (B) The pre-trained model present in MONAI-Label, trained on the VERSE
427 dataset averages a Hausdorff Distance of 3.129 ± 0.322 mm between model segmentation
428 prediction and ground truth.

429
430 **Supplementary Figure 5 – Statistical significance analysis between the quality**
431 **segmentation metrics for each method used.** (A) On average, both models, spineCT-

432 ResSegNet and spineCT-nnUNet outputs are similar with the later one outperforming the MONAI-
433 Label native model, only in regard to Hausdorff distance. Thus, both models appear to have a
434 relatively equal prediction quality. (B) The newly trained spineCT-ResSegNet model mostly
435 outperforms the pre-built spine segmentation model present in the MONAI-Label framework.
436 Further, the model was trained in one extra class, the sacrum, capable of performing robust
437 predictions for the anatomical class.

Figure 1

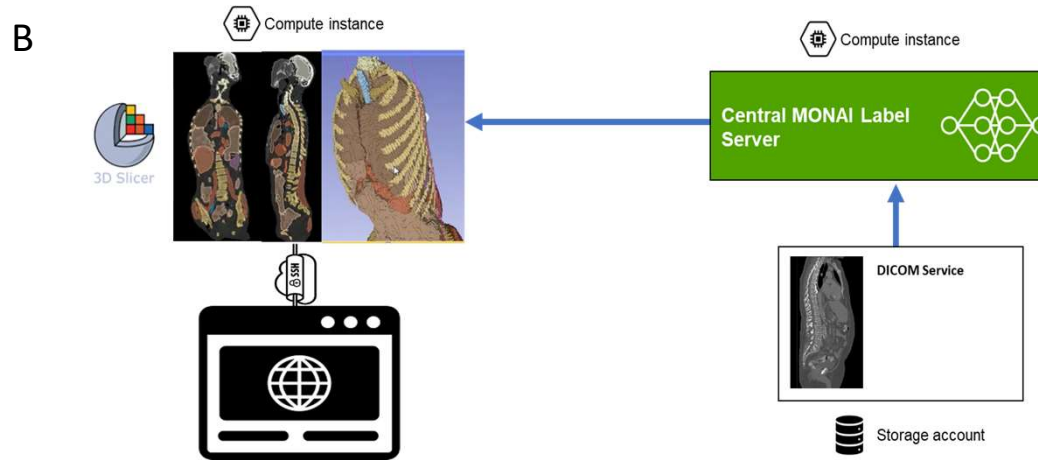
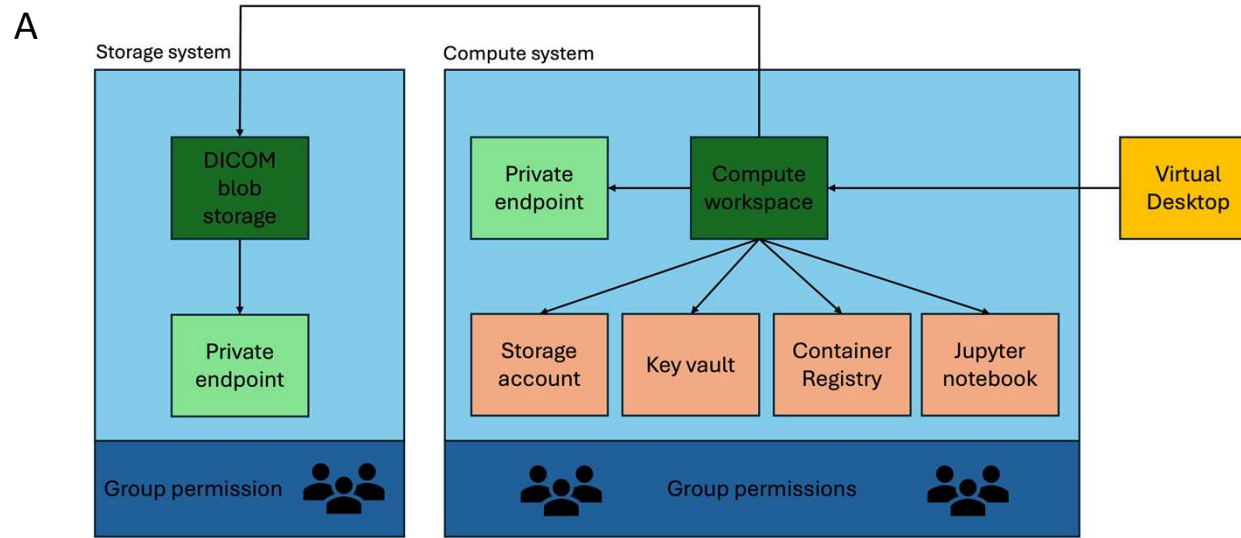
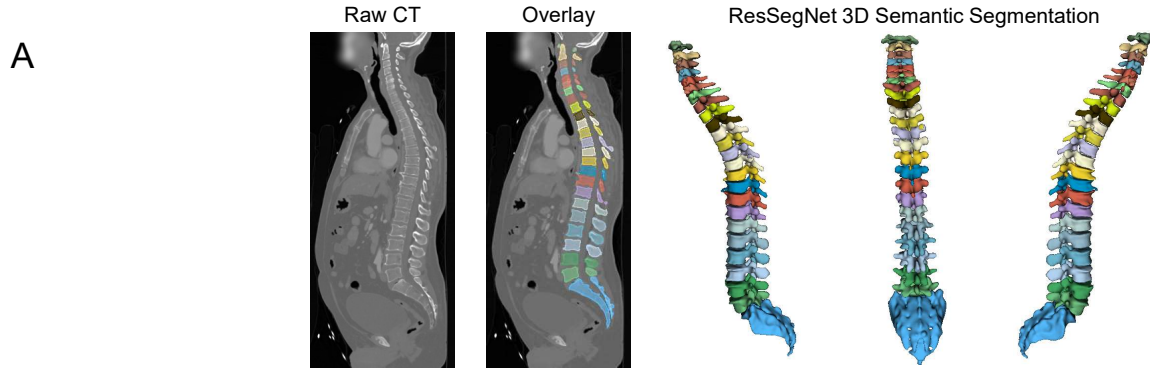


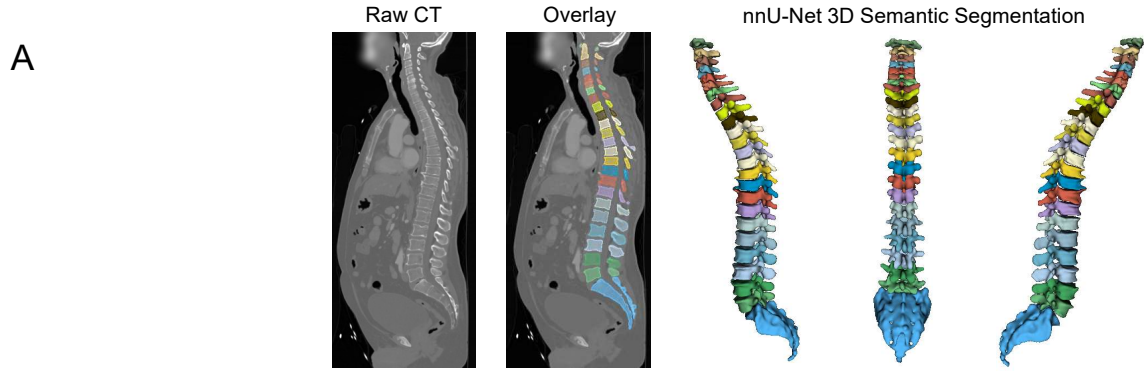
Figure 2



B

Metric	DICE Score		IoU		Accuracy		Precision		Sensitivity		Specificity		Boundary F1 Score		AUC		Kappa	
	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD
C1	0.807	0.330	0.758	0.314	1.000	0.000	0.796	0.327	0.820	0.336	1.000	0.000	0.792	0.325	0.910	0.168	0.806	0.330
C2	0.891	0.219	0.846	0.226	1.000	0.000	0.877	0.233	0.915	0.212	1.000	0.000	0.880	0.212	0.957	0.106	0.891	0.219
C3	0.891	0.211	0.840	0.207	1.000	0.000	0.907	0.204	0.877	0.213	1.000	0.000	0.894	0.134	0.939	0.107	0.890	0.211
C4	0.896	0.110	0.826	0.155	1.000	0.000	0.895	0.145	0.911	0.090	1.000	0.000	0.890	0.107	0.955	0.045	0.896	0.110
C5	0.647	0.405	0.589	0.382	1.000	0.000	0.668	0.423	0.655	0.397	1.000	0.000	0.654	0.399	0.827	0.198	0.647	0.405
C6	0.737	0.311	0.656	0.306	1.000	0.000	0.762	0.324	0.725	0.343	1.000	0.000	0.734	0.296	0.872	0.165	0.723	0.324
C7	0.812	0.222	0.724	0.225	1.000	0.000	0.874	0.204	0.792	0.219	1.000	0.000	0.783	0.196	0.896	0.110	0.812	0.222
T1	0.883	0.159	0.814	0.172	1.000	0.000	0.839	0.166	0.945	0.146	1.000	0.000	0.846	0.134	0.973	0.073	0.883	0.159
T2	0.892	0.215	0.846	0.221	1.000	0.001	0.905	0.212	0.882	0.218	1.000	0.000	0.883	0.183	0.941	0.109	0.892	0.215
T3	0.811	0.277	0.748	0.289	1.000	0.000	0.840	0.268	0.783	0.309	1.000	0.000	0.815	0.237	0.899	0.147	0.799	0.292
T4	0.804	0.282	0.739	0.295	1.000	0.001	0.775	0.289	0.850	0.268	1.000	0.000	0.787	0.251	0.925	0.134	0.803	0.282
T5	0.830	0.267	0.771	0.279	1.000	0.001	0.838	0.267	0.826	0.270	1.000	0.001	0.823	0.230	0.913	0.135	0.830	0.267
T6	0.782	0.311	0.722	0.319	1.000	0.001	0.783	0.310	0.786	0.313	1.000	0.000	0.780	0.269	0.893	0.157	0.782	0.312
T7	0.816	0.259	0.748	0.280	1.000	0.001	0.798	0.293	0.813	0.298	1.000	0.000	0.805	0.224	0.920	0.133	0.793	0.289
T8	0.879	0.195	0.819	0.214	1.000	0.001	0.858	0.234	0.888	0.207	1.000	0.000	0.850	0.173	0.951	0.089	0.866	0.219
T9	0.878	0.229	0.830	0.241	1.000	0.001	0.883	0.237	0.877	0.221	1.000	0.001	0.853	0.202	0.938	0.111	0.878	0.229
T10	0.894	0.223	0.855	0.237	1.000	0.001	0.911	0.217	0.882	0.231	1.000	0.001	0.877	0.195	0.941	0.116	0.894	0.223
T11	0.917	0.171	0.877	0.194	1.000	0.001	0.915	0.169	0.923	0.175	1.000	0.000	0.895	0.147	0.961	0.088	0.917	0.172
T12	0.915	0.166	0.872	0.190	1.000	0.001	0.909	0.195	0.909	0.194	1.000	0.000	0.885	0.152	0.961	0.081	0.903	0.195
L1	0.900	0.198	0.856	0.220	1.000	0.001	0.917	0.197	0.906	0.175	1.000	0.000	0.870	0.191	0.953	0.087	0.900	0.198
L2	0.870	0.211	0.815	0.247	1.000	0.001	0.920	0.189	0.864	0.189	1.000	0.000	0.851	0.181	0.932	0.095	0.870	0.211
L3	0.904	0.162	0.853	0.197	1.000	0.002	0.912	0.172	0.912	0.154	1.000	0.001	0.869	0.150	0.956	0.077	0.904	0.163
L4	0.883	0.211	0.833	0.230	1.000	0.001	0.908	0.218	0.889	0.177	1.000	0.001	0.846	0.188	0.944	0.088	0.883	0.211
L5	0.914	0.134	0.859	0.154	1.000	0.001	0.906	0.103	0.936	0.137	1.000	0.000	0.867	0.130	0.968	0.068	0.914	0.134
Sacrum	0.956	0.033	0.917	0.055	1.000	0.000	0.965	0.044	0.949	0.039	1.000	0.000	0.905	0.044	0.975	0.020	0.956	0.033
Average	0.856		0.801		1.000		0.862		0.861		1.000		0.837		0.932		0.853	
SD	0.066		0.074		0.000		0.066		0.070		0.000		0.057		0.034		0.068	

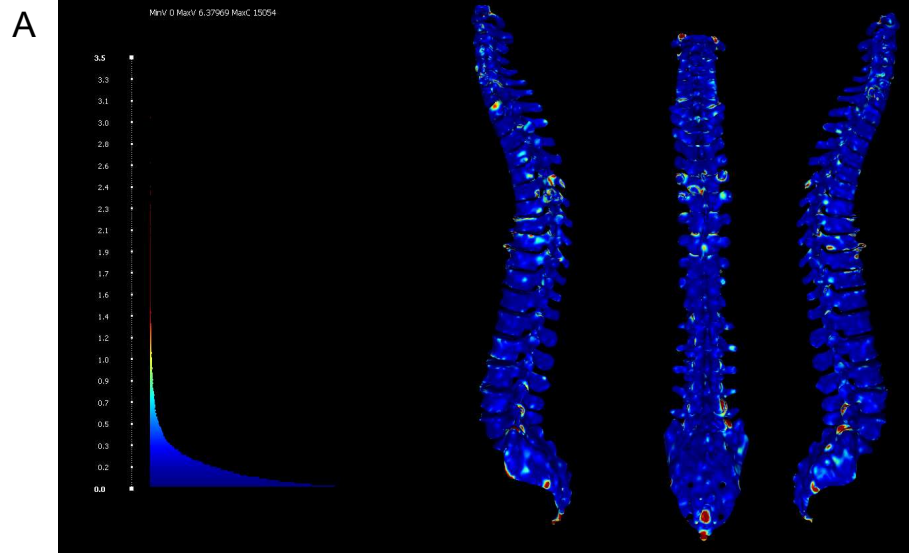
Figure 3



B

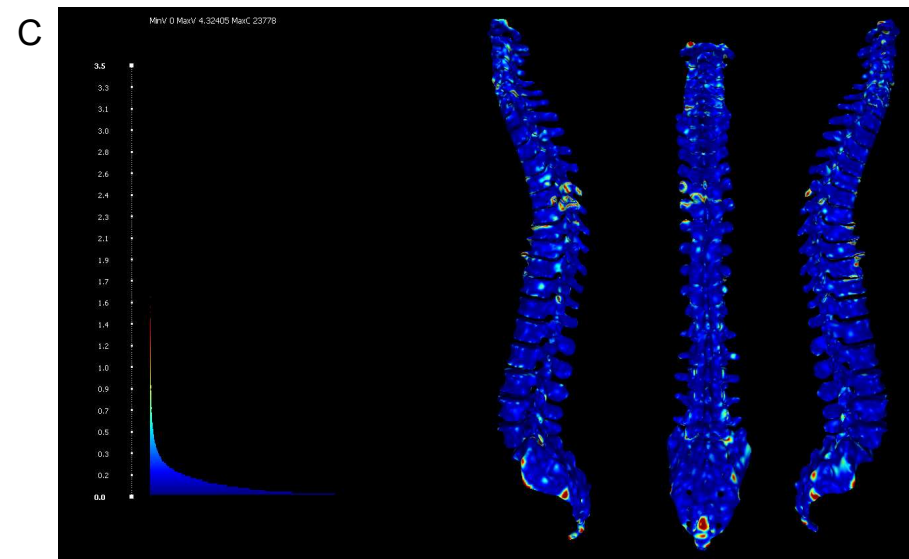
Metric	DICE Score		IoU		Accuracy		Precision		Sensitivity (recall)		Specificity		Boundary F1 Score		AUC		Kappa	
	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD
C1	0.843	0.303	0.805	0.308	1.000	0.000	0.847	0.337	0.790	0.353	1.000	0.000	0.838	0.296	0.917	0.155	0.806	0.345
C2	0.911	0.220	0.880	0.227	1.000	0.000	0.901	0.230	0.928	0.213	1.000	0.000	0.908	0.213	0.964	0.107	0.911	0.220
C3	0.936	0.086	0.890	0.123	1.000	0.000	0.970	0.025	0.914	0.122	1.000	0.000	0.943	0.033	0.957	0.061	0.936	0.086
C4	0.901	0.170	0.848	0.203	1.000	0.000	0.940	0.105	0.900	0.192	1.000	0.000	0.915	0.103	0.950	0.096	0.901	0.170
C5	0.794	0.271	0.720	0.287	1.000	0.000	0.795	0.336	0.750	0.332	1.000	0.000	0.822	0.250	0.904	0.138	0.747	0.324
C6	0.909	0.086	0.843	0.128	1.000	0.000	0.791	0.296	0.852	0.299	1.000	0.000	0.911	0.064	0.978	0.033	0.815	0.291
C7	0.923	0.148	0.879	0.164	1.000	0.000	0.907	0.170	0.959	0.080	1.000	0.000	0.907	0.132	0.979	0.040	0.923	0.148
T1	0.933	0.164	0.901	0.181	1.000	0.000	0.930	0.180	0.942	0.142	1.000	0.000	0.924	0.147	0.971	0.071	0.933	0.164
T2	0.916	0.208	0.888	0.225	1.000	0.001	0.922	0.203	0.914	0.214	1.000	0.000	0.914	0.185	0.957	0.107	0.916	0.209
T3	0.889	0.264	0.863	0.272	1.000	0.000	0.881	0.284	0.881	0.264	1.000	0.000	0.886	0.244	0.948	0.121	0.876	0.284
T4	0.878	0.274	0.850	0.279	1.000	0.000	0.881	0.281	0.893	0.243	1.000	0.000	0.872	0.255	0.946	0.121	0.878	0.274
T5	0.921	0.194	0.889	0.204	1.000	0.001	0.926	0.192	0.917	0.198	1.000	0.000	0.914	0.164	0.958	0.099	0.921	0.194
T6	0.894	0.196	0.848	0.227	1.000	0.001	0.887	0.211	0.912	0.189	1.000	0.000	0.890	0.174	0.956	0.094	0.894	0.196
T7	0.885	0.227	0.843	0.255	1.000	0.000	0.864	0.268	0.867	0.268	1.000	0.000	0.884	0.205	0.947	0.112	0.860	0.267
T8	0.934	0.154	0.900	0.177	1.000	0.000	0.935	0.154	0.936	0.159	1.000	0.000	0.920	0.141	0.968	0.079	0.933	0.154
T9	0.946	0.140	0.919	0.164	1.000	0.000	0.949	0.138	0.947	0.140	1.000	0.000	0.933	0.128	0.974	0.070	0.946	0.140
T10	0.939	0.177	0.917	0.195	1.000	0.000	0.946	0.170	0.935	0.178	1.000	0.000	0.930	0.159	0.968	0.089	0.939	0.177
T11	0.945	0.166	0.923	0.180	1.000	0.000	0.952	0.153	0.941	0.171	1.000	0.000	0.936	0.140	0.971	0.085	0.945	0.166
T12	0.935	0.181	0.910	0.200	1.000	0.000	0.921	0.211	0.933	0.201	1.000	0.000	0.926	0.152	0.973	0.085	0.922	0.209
L1	0.919	0.201	0.890	0.224	1.000	0.000	0.925	0.216	0.913	0.215	1.000	0.000	0.911	0.182	0.963	0.093	0.906	0.227
L2	0.917	0.185	0.883	0.217	1.000	0.001	0.920	0.190	0.939	0.125	1.000	0.001	0.912	0.161	0.970	0.063	0.917	0.185
L3	0.939	0.161	0.912	0.186	1.000	0.001	0.943	0.165	0.943	0.155	1.000	0.001	0.925	0.154	0.971	0.078	0.939	0.161
L4	0.926	0.199	0.902	0.218	1.000	0.001	0.936	0.180	0.932	0.193	1.000	0.000	0.918	0.163	0.966	0.097	0.926	0.199
L5	0.953	0.090	0.921	0.123	1.000	0.000	0.934	0.156	0.944	0.154	1.000	0.000	0.933	0.082	0.981	0.044	0.937	0.154
Sacrum	0.968	0.035	0.940	0.054	1.000	0.000	0.961	0.055	0.977	0.011	1.000	0.000	0.931	0.043	0.989	0.006	0.968	0.035
Average	0.914		0.879		1.000		0.911		0.910		1.000		0.908		0.961		0.904	
SD	0.036		0.045		0.000		0.046		0.050		0.000		0.029		0.018		0.049	

Figure 4



B

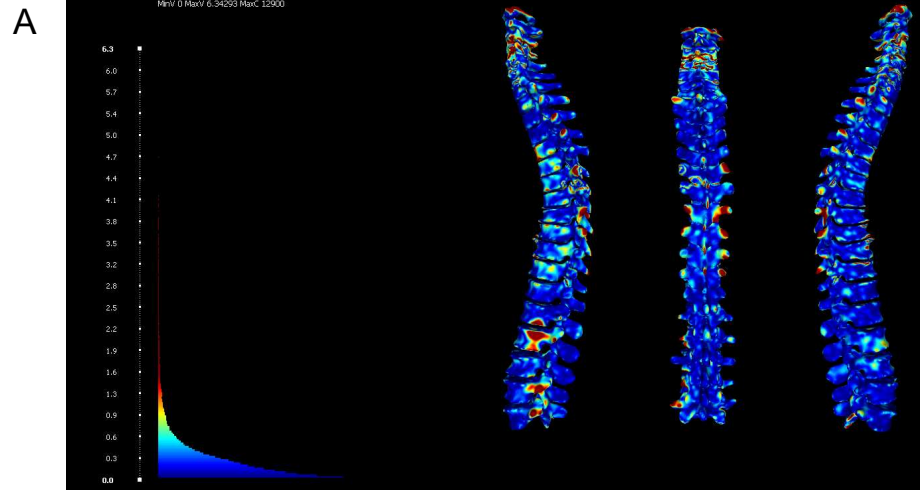
Metric Class	Hausdorff Distance (mm)		Lower 95% Confidence Interval	Upper 95% Confidence
	Average	SD		
C1	1.858	0.438	1.675	2.041
C2	2.028	0.428	1.845	2.211
C3	2.041	0.465	1.832	2.249
C4	2.096	0.655	1.809	2.383
C5	1.860	0.678	1.626	2.095
C6	1.854	0.508	1.716	1.993
C7	2.335	0.585	2.193	2.476
T1	2.639	0.585	2.498	2.780
T2	2.280	0.592	2.137	2.423
T3	2.675	0.766	2.491	2.858
T4	2.863	0.796	2.670	3.055
T5	2.743	0.843	2.533	2.953
T6	2.767	0.881	2.552	2.983
T7	2.687	0.865	2.484	2.890
T8	2.757	0.877	2.555	2.960
T9	2.650	0.791	2.471	2.829
T10	2.569	0.734	2.403	2.735
T11	2.643	0.775	2.471	2.815
T12	2.825	0.716	2.664	2.986
L1	2.914	0.793	2.730	3.099
L2	3.045	0.870	2.832	3.258
L3	3.263	0.789	3.060	3.466
L4	3.155	0.600	3.002	3.308
L5	2.983	0.548	2.841	3.125
Sacrum	3.566	0.572	3.415	3.717
Average	2.604			
SD	0.450			



D

Metric Class	Hausdorff Distance (mm)		Lower 95% Confidence Interval	Upper 95% Confidence
	Average	SD		
C1	1.646	0.402	1.478	1.814
C2	1.753	0.474	1.550	1.955
C3	1.861	0.485	1.643	2.079
C4	1.995	0.702	1.687	2.303
C5	1.706	0.738	1.450	1.962
C6	1.609	0.472	1.481	1.737
C7	1.855	0.508	1.733	1.978
T1	2.022	0.691	1.856	2.189
T2	2.015	0.634	1.862	2.168
T3	2.130	0.652	1.974	2.286
T4	2.215	0.726	2.040	2.390
T5	2.307	0.858	2.097	2.517
T6	2.140	0.871	1.936	2.344
T7	2.173	0.927	1.959	2.387
T8	2.219	0.837	2.029	2.408
T9	2.177	0.684	2.022	2.332
T10	2.284	0.755	2.117	2.452
T11	2.475	0.811	2.293	2.658
T12	2.547	0.945	2.327	2.766
L1	2.602	1.014	2.354	2.850
L2	2.725	0.832	2.511	2.939
L3	2.642	0.800	2.438	2.846
L4	2.507	0.666	2.334	2.680
L5	3.452	0.571	3.301	3.603
Sacrum	2.287	0.816	2.084	2.490
Average	2.214			
SD	0.398			

Supplementary Figure 4



B

Metric Class	Hausdorff Distance (mm)		Lower 95% Confidence Interval	Upper 95% Confidence Interval
	Average	SD		
C1	2.820	0.875	2.454	3.185
C2	2.965	0.647	2.688	3.242
C3	2.782	0.576	2.523	3.041
C4	2.918	0.713	2.606	3.231
C5	2.370	0.954	2.040	2.701
C6	2.482	0.652	2.305	2.659
C7	3.100	0.773	2.913	3.286
T1	3.373	0.879	3.161	3.585
T2	2.785	0.667	2.624	2.946
T3	3.032	0.844	2.830	3.234
T4	3.107	0.794	2.915	3.298
T5	3.348	0.757	3.160	3.537
T6	3.532	0.887	3.315	3.749
T7	3.479	0.946	3.258	3.701
T8	3.508	0.838	3.315	3.702
T9	3.331	0.907	3.126	3.536
T10	3.156	1.027	2.924	3.388
T11	2.972	0.919	2.768	3.176
T12	3.077	0.935	2.867	3.287
L1	3.146	0.936	2.928	3.364
L2	3.340	0.999	3.096	3.585
L3	3.622	1.092	3.341	3.903
L4	3.479	1.024	3.218	3.740
L5	3.376	0.921	3.137	3.615
Average	3.129			
SD	0.322			

Supplementary Figure 5

A

Metric	p-value spineCT-ResNetSeg vs spineCT_nnUNet									
Class	Dice Score	IoU	Accuracy	Precision	Sensitivity	Specificity	Boundary F1 Score	AUC	Kappa	Hausdorff Distance
C1	0.70459	0.62022	0.37299	0.60639	0.77369	0.23583	0.63048	0.88678	0.99676	0.10134
C2	0.76732	0.62264	0.04599	0.74846	0.84630	0.07175	0.67279	0.84609	0.76712	0.05541
C3	0.38920	0.37530	0.81518	0.19540	0.52399	0.08493	0.13669	0.52379	0.38927	0.25019
C4	0.91825	0.69662	0.70715	0.26210	0.82387	0.11154	0.45383	0.82425	0.91819	0.64195
C5	0.08377	0.12542	0.98932	0.18394	0.29629	0.24132	0.04939	0.07429	0.27292	0.38644
C6	0.00029	0.00013	0.55940	0.62289	0.03974	0.28055	0.00009	0.00002	0.11968	0.01228
C7	0.00100	0.00002	0.01311	0.32447	0.00000	0.70594	0.00004	0.00000	0.00099	0.00000
T1	0.07743	0.00520	0.29485	0.00300	0.90277	0.20333	0.00176	0.90530	0.07739	0.00000
T2	0.50221	0.28602	0.77267	0.63746	0.38511	0.82585	0.32729	0.38542	0.50240	0.01414
T3	0.09918	0.01880	0.43897	0.38910	0.05015	0.69282	0.08713	0.03676	0.12344	0.00002
T4	0.12399	0.02828	0.20584	0.03377	0.34078	0.21254	0.05540	0.34000	0.12387	0.00000
T5	0.03212	0.00940	0.33326	0.33251	0.03608	0.29437	0.01227	0.03611	0.03218	0.00273
T6	0.01649	0.01140	0.37949	0.02841	0.00710	0.48161	0.00721	0.00712	0.01653	0.00330
T7	0.09539	0.03637	0.30951	0.16098	0.25788	0.11183	0.03190	0.19187	0.14948	0.00028
T8	0.06248	0.01513	0.12454	0.01994	0.12454	0.08958	0.00857	0.23178	0.03474	0.00016
T9	0.02940	0.00959	0.13170	0.03971	0.02133	0.15602	0.00441	0.02125	0.02933	0.00145
T10	0.16943	0.07827	0.17091	0.27567	0.11463	0.24472	0.06896	0.11449	0.16902	0.00091
T11	0.30885	0.12574	0.33137	0.14648	0.50813	0.30453	0.07663	0.50723	0.30843	0.00396
T12	0.49914	0.23033	0.52564	0.69929	0.46163	0.98056	0.09686	0.38318	0.56405	0.00652
L1	0.57987	0.36998	0.78695	0.82605	0.83668	0.95113	0.19787	0.49408	0.86899	0.01320
L2	0.18699	0.10461	0.68086	0.98805	0.00890	0.76718	0.04443	0.00897	0.18701	0.00902
L3	0.24919	0.10319	0.47364	0.32719	0.27620	0.44883	0.05272	0.27548	0.24888	0.00051
L4	0.25326	0.09455	0.45636	0.43427	0.20754	0.33812	0.02724	0.20696	0.25294	0.00015
L5	0.06874	0.02030	0.44177	0.25953	0.76385	0.17086	0.00164	0.24012	0.39430	0.00006
Sacrum	0.06283	0.03119	0.15076	0.69608	0.00000	0.64117	0.00198	0.00000	0.06199	0.29829

B

Metric	p-value spineCT-ResNetSeg vs MONAI-Verse									
Class	Dice Score	IoU	Accuracy	Precision	Sensitivity	Specificity	Boundary F1 Score	AUC	Kappa	Hausdorff Distance
C1	0.00009	0.00000	0.00999	0.26757	0.00000	0.04576	0.00014	0.00000	0.00009	0.00007
C2	0.01107	0.00219	0.00194	0.03920	0.00309	0.03918	0.00393	0.00394	0.01103	0.00000
C3	0.00162	0.00011	0.00301	0.21819	0.00011	0.77313	0.00054	0.00008	0.00162	0.00011
C4	0.00031	0.00005	0.00841	0.03238	0.00001	0.29519	0.00022	0.00001	0.00031	0.00051
C5	0.03262	0.01585	0.04449	0.02216	0.03347	0.01699	0.02030	0.03712	0.03256	0.01675
C6	0.00000	0.00000	0.01378	0.00000	0.00000	0.28034	0.00000	0.00000	0.00000	0.00000
C7	0.00000	0.00000	0.00285	0.00000	0.00000	0.64321	0.00000	0.00000	0.00000	0.00000
T1	0.00000	0.00000	0.04222	0.75315	0.00000	0.04798	0.00000	0.00000	0.00000	0.00000
T2	0.03197	0.00472	0.49916	0.04358	0.03495	0.60738	0.00550	0.06750	0.03181	0.00001
T3	0.04069	0.03442	0.10117	0.01611	0.26632	0.06077	0.01945	0.15439	0.08519	0.01131
T4	0.00098	0.00093	0.17255	0.10940	0.00001	0.41258	0.00100	0.00000	0.00098	0.08011
T5	0.00004	0.00001	0.33749	0.00890	0.00000	0.70684	0.00001	0.00000	0.00004	0.00005
T6	0.00000	0.00000	0.16731	0.00313	0.00000	0.68572	0.00000	0.00000	0.00000	0.00000
T7	0.00000	0.00000	0.16460	0.00021	0.00000	0.97582	0.00000	0.00000	0.00000	0.00000
T8	0.00000	0.00000	0.11214	0.00102	0.00000	0.96214	0.00000	0.00000	0.00000	0.00000
T9	0.00000	0.00000	0.25478	0.02706	0.00000	0.87060	0.00000	0.00000	0.00000	0.00000
T10	0.00080	0.00026	0.35119	0.00841	0.00013	0.78478	0.00011	0.00014	0.00081	0.00009
T11	0.00752	0.00576	0.36233	0.14362	0.00040	0.65906	0.00195	0.00094	0.00414	0.01684
T12	0.00295	0.00248	0.14787	0.62042	0.00101	0.94077	0.00269	0.00051	0.00541	0.06433
L1	0.00221	0.00150	0.22585	0.07469	0.00034	0.44340	0.00194	0.00035	0.00220	0.11389
L2	0.00509	0.00474	0.46612	0.02918	0.00089	0.52267	0.00107	0.00095	0.00511	0.07658
L3	0.00171	0.00157	0.57122	0.04284	0.00006	0.74193	0.00083	0.00033	0.00051	0.04490
L4	0.00731	0.00483	0.35114	0.01723	0.00127	0.41694	0.00448	0.00130	0.00731	0.03846
L5	0.00012	0.00004	0.11469	0.01287	0.00000	0.27333	0.00011	0.00000	0.00006	0.00686