Good Enough? An Investigation on the Impact of Label Quality in Large-Scale Medical Datasets

Alexander Jaus[1,2*], Zdravko Marinov[1,2], Simon Reiß[1], Constantin Seibold[3], Jiale Wei[1,2], Jens Kleesiek[4†], and Rainer Stiefelhagen[1†]

[1] Karlsruhe Institute of Technology, Germany
{firstname.lastname}@kit.edu
[2] HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany
[3] Diagnostic and Interventional Radiology, Heidelberg University Hospital
{firstname.lastname}@med.uni-heidelberg.de
[4] Institute for AI in Medicine, University Hospital Essen, Essen, Germany
{firstname.lastname}@uk-essen.de

Abstract. Manually refining radiological segmentation masks is highly resource-intensive. To determine when this expert commitment is truly justified for the training of segmentation models, we investigate the relationship between label quality and model performance. Expanding beyond models trained directly for inference, we conduct the first study isolating the impact of label quality in pre-training datasets. While high-quality labels remain essential for models proceeding directly to deployment, we find no evidence that strict label quality is crucial for pre-training efficacy. These results question the necessity of exhaustive human-in-the-loop refinement for massive corpora intended for pretraining and suggest that expert effort is more effectively invested in well-curated downstream target datasets.

Keywords: Pretraining · Anatomy Segmentation · Pseudolabels

1 Introduction

Annotating a dataset the size of the recently proposed DAP-Atlas [8] would require a radiologist to spend over 12,000 hours, assuming an optimistic 10 minutes per mask. Other large-scale datasets [27, 20, 13] are no exception. These massive time requirements make manual annotation clearly infeasible at scale, pushing researchers toward alternative strategies to ensure high-quality masks: DAP-Atlas [8] relies on algorithmic checks and includes radiologists only for validation, TotalSegmentator [27] employs a human-in-the-loop refinement, and

*Corresponding author (E-mail: alexander.jaus@kit.edu)
† indicates equal supervision

AbdomenAtlas [20, 13] uses uncertainty-based guidance to identify automatically generated masks in need of corrections. This marks a shift in how large datasets are created, where the common approach has shifted from manually annotating masks to refining automatically generated masks. While this significantly reduces manual workload, it still requires substantial radiologist involvement, with dataset curation still taking weeks of radiologist involvement [20] while leaving label quality uncertain [12] due to sparse verification and the assumption that model uncertainty perfectly correlates with actual error. This raises a critical research question: How do these inherent annotation errors influence the performance of segmentation models trained on such labels?

To address this question, we simulate the construction of modern large-scale datasets by replacing high-quality ground-truth labels with predictions from a diverse suite of models, including nnU-Net [5] and several medical foundation models [16, 27, 4]. Comparing predictions to the actual ground truth, we generate dataset variants with varying levels of label quality. We then train segmentation models on these "noisy" versions, thereby mimicking a scenario where researchers unknowingly train on imperfect, automatically generated labels. These models are evaluated across two common scenarios: In-dataset performance against the dataset's clean base version and pretraining on the noisy data followed by fine-tuning on an independent, high-quality dataset. These tasks reflect two standard workflows: direct deployment after training on noisy data, or large-scale pre-training followed by high-quality downstream fine-tuning. We find a significant disparity between these tasks: while label quality is critical for models used directly after training, fine-tuning largely compensates for poor-quality pretraining labels. Since superior pretraining labels do not consistently yield better fine-tuned models, we hypothesize that during pretraining, general structural knowledge is obtained rather than details.

Contributions: (1) We conduct a detailed study of label noise in medical segmentation across four datasets and seven labeling models, using realistic pseudo-label noise mimicking modern automated curation pipelines. (2) We perform, to the best of our knowledge, the first benchmark isolating the relationship between pre-training label quality and final downstream performance after fine-tuning. (3) We reveal a critical dichotomy: while high label quality is essential for in-domain deployment, fine-tuning effectively compensates for noisy pre-training labels. (4) We derive actionable insights for dataset creators and model developers.

Related Work: Our study contributes to the field of label-noise [23], where most of the related works come from the task of building robust models via the estimation of noise-confusion-matrices [22, 25], regularization [7], or robust losses [2]. While valuable, these studies typically aim to adapt models to noise and compare results against non-adapted baselines. In contrast, we observe that researchers often train standard models on newly released datasets without explicit noise correction. We therefore fix the model architecture and systematically vary the data to emulate this realistic setting of unknown label quality. Furthermore, existing studies on label-noise impact often focus on RGB modalities [28],

microscopy [10], or utilize 2D models for volumetric data [3, 19, 26], thereby ignoring critical inter-slice relationships. The works which natively work with 3D data often focus on single, very specific structures, such as the mandibles [30] or the parotid glands [24], and work on small application-specific datasets.
Existing studies share a significant limitation: noise is typically introduced synthetically via morphological operations [1, 24] or perturbations [3], which bake prior assumptions into the error model. Others focus on inter-annotator disagreement [28, 30], a setting ideal for manually segmented datasets, but falling short of capturing the construction pipelines of modern datasets [8, 20, 27] which heavily rely on pseudolabels. We address these gaps by modeling noise as realistic pseudo-label errors across datasets of varying scales. Furthermore, we utilize native 3D models to preserve volumetric context and provide the first analysis of how label quality specifically impacts the pre-training phase.

2 Methodology

2.1 From Labels to Pseudo-Labels

Pseudo-labeled datasets are defined as derivatives of a base dataset D as

$$D_g = \{(X^i, g(X^i))\}_{i=1}^N \tag{1}$$

where g is a neural network, capable of generating predictions $g(X^i) = \hat{Y}_g^i$ based on the input X^i. Within this work, we consider a set of pseudo-label generators $\mathcal{G}$ ={nnU-Net [5], MedSAM [16], STU-Net$_{\{\text{small, base, large, huge}\}}$ [4], TotalSegmentator [27]}, allowing us to generate a total of 7 different pseudo-label version (one for each $g \in \mathcal{G}$) per base dataset. We define $\mathcal{G}^+ = \mathcal{G} \cup \{\text{base}\}$ to include the original dataset labels.
Pseudo-Label Generators: We employ three categories of generators, each representing a distinct family of medical segmentation models. The first category comprises standard trainable networks, represented by nnU-Net [5], being one of the most popular and successful segmentation frameworks. We use five-fold cross-validation, training separate models for each fold of the base datasets D_{base}. Each model is trained on four folds and predicts the respective fifth fold, which we keep as pseudolabels. Aggregating these out-of-fold predictions yields $D_{\text{nnU-Net}}$, now composed entirely of predicted labels.

The second category comprises "out-of-the-box" foundation models: TotalSegmentator [27] and the STU-Net family [4]. TotalSegmentator is a widely used segmentation model that is capable of robustly segmenting large parts of the human anatomy. STU-Net is a family of U-Net [21] based networks which differ in parameter sizes. The authors find that scaling model parameters tends to slightly but consistently improve segmentation results. For our study, this poses an opportunity, as the small differences in performance that the different STU-Net variants yield are an ideal simulation of small quality fluctuations of annotations; this behavior is explored in Figure 1 for the Word [15] dataset, with other datasets behaving similarly.

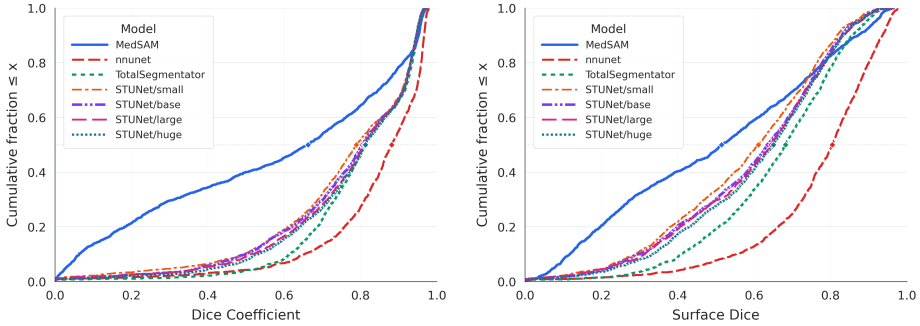


Fig. 1: Empirical cumulative distribution functions (ECDFs) of pseudo-label quality for base dataset Word. Each curve shows the proportion of predictions whose dice/surface-dice value is at most x, averaged over all anatomical classes. A curve shifted further to the right indicates higher overall segmentation quality. nnU-Net pseudo-labels consistently dominate, while MedSAM exhibits the lowest quality. Intermediate positions are rather close but are typically ordered from STUNet, small to huge, with TotalSegmentator in between.

Finally, we use the interactive foundation model MedSAM [16] serving as a bridge between the in-domain dataset generators being perfectly able to capture the annotation style of the dataset, and non-interactive models whose predictions cannot be altered. For MedSAM, we slice volumetric images and generate bounding boxes around ground-truth segmentation masks to prompt the model, allowing it to adapt to the original labels without requiring training.

Selection of base Datasets: As base datasets D, we select WORD [15] (100 cases, 16 organs), AMOS [9] (200 cases, 15 organs), CT-1K [18] (1,000 cases, 4 organs), and AbdomenAtlas [20, 13] (5,200 cases, 9 organs). These abdominal datasets were selected to represent a range of popular benchmarks and scales, facilitating a systematic study of label quality across varying organs and data magnitudes.

2.2 Assessing the Quality of the Pseudo-Label datasets

For each base dataset D, we compute the different pseudo-label variants as outlined in Equation (1). Following [18], we measure label quality via Dice and Surface Dice to assess how well each pseudo-label dataset (e.g., $Word_{\text{nnUnet}}$) matches its original labels ($Word_{\text{base}}$). We report the results in Table 1 by averaging first across the classes, followed by averaging across samples for both Dice and Surface Dice. To aggregate the organ-level standard deviations, we compute the root-mean-square across the images in the dataset. Ensuring a fair comparison, we only consider anatomical structures that all generators $g \in \mathcal{G}$ are capable of predicting and ignore the others in all averages. This is necessary since the non-interactive pseudo-label predictors only offer a fixed set of labels that cannot be altered. These are, however, only minimal adjustments affecting

Table 1: Overview of the quality of different pseudo-label dataset variants

		Word	Amos	CT1K	Abd. Atlas
nnUnet	Dice	83.7±12.8	89.2±12.6	95.2±4.3	90.6±13.7
	Surf. Dice	76.5±8.7	85.2±9.2	88.9±9.4	84.4±9.2
MedSAM	Dice	56.8±18.6	59.9±23.7	72.9±13.6	39.4±20.5
	Surf. Dice	49.6±7.0	51.9±7.2	57.8±7.6	24.8±5.0
TotalS	Dice	79.3±10.3	82.9±13.7	91.9±5.5	86.6±16.4
	Surf. Dice	65.4±8.1	71.6±8.5	80.3±9.0	76.1±8.7
STU S	Dice	74.8±13.6	80.9±15.5	91.5±6.5	83.2±19.9
	Surf. Dice	57.6±7.6	66.8±8.2	78.6±8.9	70.0±8.4
STU B	Dice	75.9±12.5	82.1±15.5	92.0±6.4	84.7±18.3
	Surf. Dice	59.5±7.7	66.7±8.3	80.1±8.9	72.7±8.5
STU L	Dice	76.5±12.9	82.7±15.1	92.2±6.0	85.2±18.2
	Surf. Dice	60.0±7.7	69.6±8.3	80.6±9.0	73.8±8.6
STU H	Dice	77.2±12.1	83.0±14.6	92.2±5.8	85.8±17.4
	Surf. Dice	60.9±7.8	69.7±8.3	80.8±9.0	74.4±8.6

the class *Prostate* in Amos and *Rectum* in Word, which are excluded from the averages alongside the *Background* class.

Upon inspection of the results in Table 1 and Figure 1, we report the following findings: (1) There is a clear improvement pattern from STU S $\rightarrow$ B $\rightarrow$ L $\rightarrow$ H, though the gains between models become smaller with increasing model size, confirming the findings of [4]. We show this behavior in more detail in Figure 1 for the Word dataset, with the other base datasets behaving similarly. (2) nnUNet outperforms on most datasets, achieving Dice scores from 83.7% to 95.2%, benefiting from direct training on each dataset's annotation style. (3) MedSAM, as the candidate for an interactive foundation model, generally has a lower quality of predictions, which can be attributed to the model natively operating in 2D, making it unable to capture the volumetric nature of the data. Stacking its 2D predictions back into 3D introduces artifacts besides the in-slice errors. While there are approaches to adapt SAM-based models to 3D (e.g. via adapters [29]), this is not the goal of this work, and we deliberately treat the reduced label quality of MedSAM's predictions as a proxy for low-quality slice-based labels, thereby spanning a realistic spectrum of pseudo-label quality from lower to higher regimes.

3 Experiments and Results

We evaluate the impact of dataset quality across two scenarios: In-domain evaluation and pre-training suitability.

Training Setup: We chose DynUnet [6] as our model of choice, as it offers a flexible and well-proven architecture serving as the workhorse of the nnU-

Net [5] framework. To mitigate any influence on model training except for the data, we use a completely deterministic training for 1000 epochs, where each epoch consists of exactly 250 mini-batches, thereby oversampling the dataset if needed. We generally follow the nnUnet training setup by resampling images to their respective datasets' median resolution before extracting patches. For the in-domain evaluation setup, we choose patch sizes to be specified by each dataset D and remain constant across the different versions $D_g \forall g \in \mathcal{G}^+$. The network weights are updated using the AdamW [14] optimizer over a cosine-annealing learning rate scheduler starting from a base learning rate of $1e-3$. We leverage random 80/20 splits, setting aside 80% of the data for training and 20% for evaluation. The splits are consistent across all versions of D.

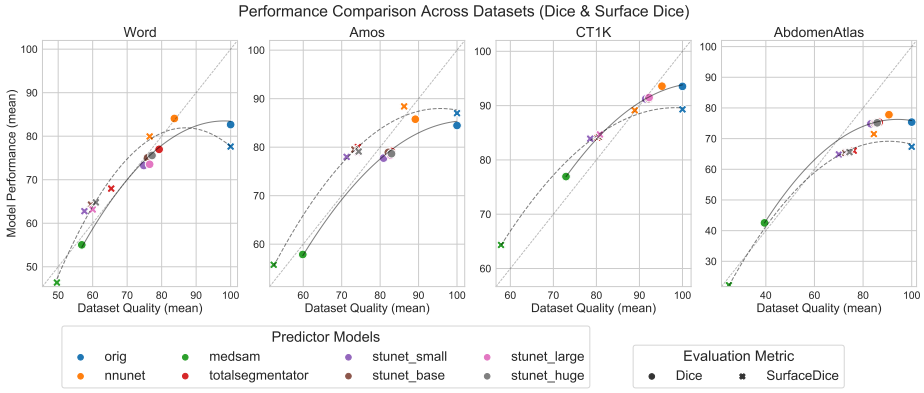


Fig. 2: In-Domain evaluation results for Dice and Surface Dice. We include dashed $y = x$ for reference and perform a second degree polynomial regression on the Dice data (solid line) and Surface Dice (dashed line).

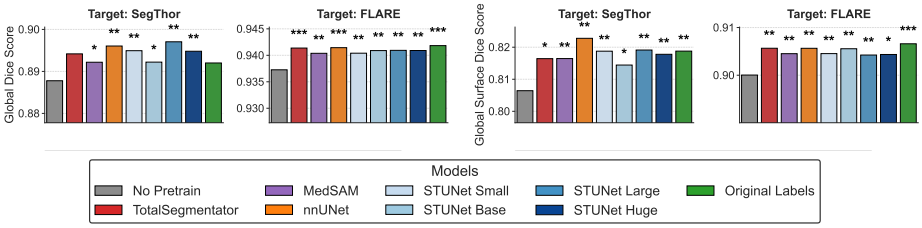
3.1 In-Domain Evaluation

For the task of in-domain evaluation, we consider the setup, where a model $f_{D_g}, g \in \mathcal{G}^+$, that was trained on D_g is evaluated against the testset of D_{base}. This scenario reflects the setting in which the target structures are predefined, yet no manual ground-truth annotations are accessible for training, such that models may be trained solely on pseudo-labels and are directly evaluated on the original reference distribution. We report the results in Figure 2 and summarize our findings as follows: (1) Overall, we see a strong positive correlation between dataset quality and model performance. As the quality of the dataset increases, there is a distinct rise in the resulting model performance. (2) The relationship between label quality and performance is non-linear, showing diminishing returns at the higher end of the spectrum. As the dataset quality approaches 100, the performance curves begin to plateau, falling below the diagonal reference

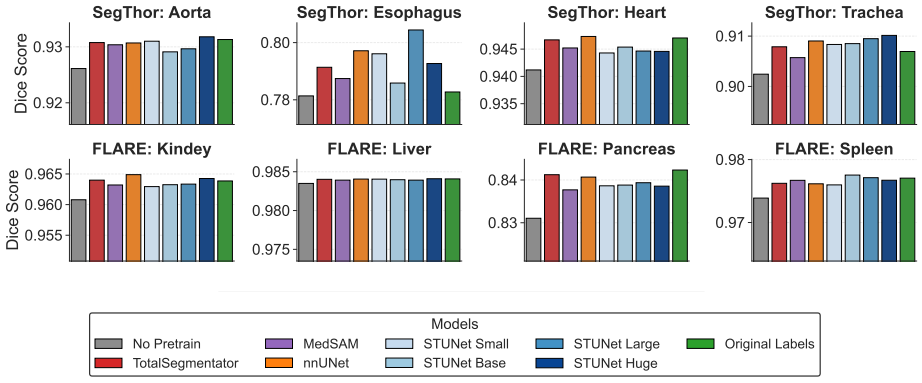
line. This indicates that minor improvements in already high-quality labels do not necessarily translate to proportional performance gains. (3) We observe a scaling effect where dataset volume appears to compensate for label quality. In smaller datasets such as Word or Amos, improving labels of already high quality is still reflected in considerable gains, whereas in the massive AbdomenAtlas dataset (5k cases), performance plateaus early, suggesting that sheer data quantity compensates for imperfect labels. Consequently, the "quality-performance" correlation is most critical for smaller cohorts or low-quality datasets. (4) Among the intermediate pseudo-label generators (STU-Net variants and TotalSegmentator), incremental improvements in label quality do not guarantee monotonic performance gains, as minor enhancements may not provide enough signal to make a clear difference. (5) Boundary accuracy (Surface Dice) consistently lags behind volumetric overlap (Dice) across all quality levels, indicating that precise surface segmentation remains more challenging than general overlap.

3.2 Pre-training Suitability

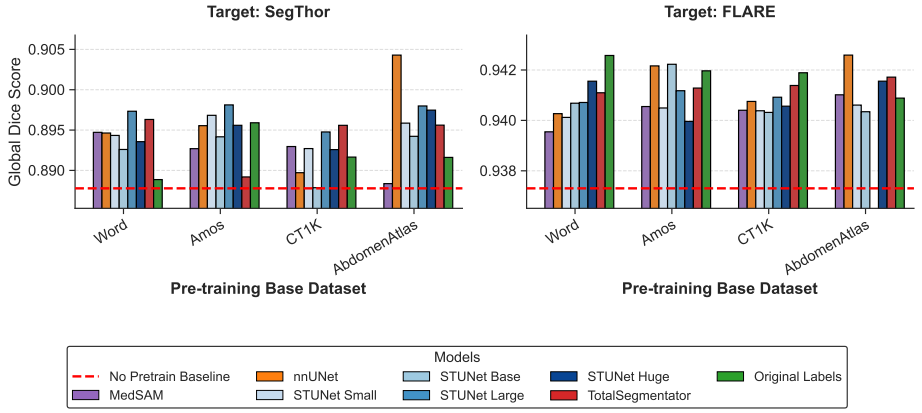
We now turn towards the scenario of pre-training models on the generated noisy dataset before fine-tuning them on one of two different clean target datasets: SegThor [11] and Flare [17]. While the core training procedure remains consistent, we adapt the patch size and normalization schemes to match the target datasets. For each pre-trained model $f_{D_g}, g \in \mathcal{G}^+$, we initialize the fine-tuning stage with the pre-trained weights and repeat the training process on the target data. The resulting models are evaluated on the respective test sets of the target datasets, with findings detailed across three subplots in Figure 3. In the first subplot, we present global averaged scores aggregated across all organs and base datasets (e.g., Word, Amos,...). We find that for the vast majority of pretraining settings, we outperform a non-pretrained baseline significantly, as indicated by a one-sided Wilcoxon signed-rank test. We verified these gains are data-driven, as extending the baseline training from 1k to 2k steps (matching the total pretraining and fine-tuning duration) yielded no performance improvement. More importantly, we find that the label quality of the pre-training datasets no longer seems to be of concern. This is most evident in the MedSAM results: despite the poor label quality that hindered the in-domain performance of $f_{D_{MedSAM}}$ (Figure 2), MedSAM remains highly effective as a pre-training source. Models trained on MedSAM labels significantly outperform the non-pretrained baseline and achieve scores comparable to models pretrained on datasets with substantially higher label quality. This observation is further supported by the organ-level scores in subplot (b). While certain structures, such as the pancreas and trachea, exhibit greater gains from pre-training compared to other labels, such as the liver, these improvements are not primarily driven by the label quality of the pre-training dataset. Across diverse anatomical structures, the performance boosts remains consistent regardless of whether the pre-training labels were high-quality or noisy. Subfigure (c) details the effect of pre-training dataset selection on downstream performance. Interestingly, under a fixed compute budget (the



(a) Dice and Surface Dice scores on the SegThor and FLARE benchmarks. Each bar represents the average scores across different base datasets and individual organs. Stars above a bar indicate that the model significantly outperforms the *No Pretrain* setting according to a one-sided Wilcoxon signed-rank test (*: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$)



(b) Dice scores of individual organ masks on SegThor and Flare. Each bar represents the averaged organ-mask scores across the different base-datasets.



(c) Dice scores of averaged organ mask on SegThor and Flare datasets, highlighting details on the effects of the pre-training base-datasets and label-generators.

Fig. 3: Pre-training evaluation results: we report global and organ-specific segmentation performances on the SegThor and FLARE datasets, alongside details regarding the relationship of base datasets and label generators.

default for nnU-Net style training), dataset size does not appear to have a major impact. Models pre-trained on Amos or Word, which are significantly smaller than AbdomenAtlas, do not seem to learn less meaningful representations. Training runs where fine-tuning did not converge under the described deterministic procedure are indicated by omitted bars.

4 Conclusion

We present the first large-scale study isolating the impact of label quality and dataset size across medical image segmentation paradigms. Our results show that while high-quality annotations are essential for models deployed directly after training, these requirements largely disappear in pretraining settings. Fine-tuning compensates for noisy pretraining labels, suggesting that pretraining primarily transfers general structural knowledge rather than precise details. We derive two actionable insights: (i) For dataset creators targeting massive pretraining datasets, extensive expert refinement may not be cost-effective. (ii) For model developers, simple pseudo-labeling of unlabeled image collections can suffice to learn transferable representations.

Acknowledgments. The present contribution is supported by the Helmholtz Association under the joint research school "HIDSS4Health - Helmholtz Information and Data Science School for Health. This work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research. This work was supported by funding from the pilot program Core-Informatics of the Helmholtz Association (HGF).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Brückner, C., Liu, C., Rist, L., Maier, A.: Influence of imperfect annotations on deep learning segmentation models. In: BVM Workshop. pp. 226–231. Springer (2024)
2. Gonzalez-Jimenez, A., Lionetti, S., Gottfrois, P., Gröger, F., Navarini, A., Pouly, M.: Robust t-loss for medical image segmentation. Medical Image Analysis p. 103735 (2025)
3. Heller, N., Dean, J., Papanikolopoulos, N.: Imperfect segmentation labels: How much do they matter? In: International Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis. pp. 112–120. Springer (2018)
4. Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., et al.: Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716 (2023)
5. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

6. Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128 (2019)
7. Islam, M., Glocker, B.: Spatially varying label smoothing: Capturing uncertainty from expert annotations. In: international conference on information processing in medical imaging. pp. 677–688. Springer (2021)
8. Jaus, A., Seibold, C., Hermann, K., Shahamiri, N., Walter, A., Giske, K., Haubold, J., Kleesiek, J., Stiefelhagen, R.: Towards unifying anatomy segmentation: Automated generation of a full-body ct dataset. In: 2024 IEEE International Conference on Image Processing (ICIP). pp. 41–47. IEEE (2024)
9. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)
10. Jiménez, L.G., Decaestecker, C.: Impact of imperfect annotations on cnn training and performance for instance segmentation and classification in digital pathology. Computers in biology and medicine **177**, 108586 (2024)
11. Lambert, Z., Petitjean, C., Dubray, B., Kuan, S.: Segthor: Segmentation of thoracic organs at risk in ct images. In: 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6. IEEE (2020)
12. Li, W., Qu, C., Chen, X., Bassi, P.R., Shi, Y., Lai, Y., Yu, Q., Xue, H., Chen, Y., Lin, X., et al.: Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. Medical Image Analysis **97**, 103285 (2024)
13. Li, W., Yuille, A., Zhou, Z.: How well do supervised 3d models transfer to medical imaging tasks? In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=AhizIPytk4
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
15. Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S.: Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. Medical Image Analysis **82**, 102642 (2022)
16. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
17. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., Gou, S., Thaler, F., Payer, C., Štern, D., Henderson, E.G., McSweeney, D.M., Green, A., Jackson, P., McIntosh, L., Nguyen, Q.C., Qayyum, A., Conze, P.H., Huang, Z., Zhou, Z., Fan, D.P., Xiong, H., Dong, G., Zhu, Q., He, J., Yang, X.: Fast and low-gpu-memory abdomen ct organ segmentation: The flare challenge. Medical Image Analysis **82**, 102616 (2022)
18. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2021)
19. Marcinkiewicz, M., Mrukwa, G.: Quantitative impact of label noise on the quality of segmentation of brain tumors on mri scans. In: 2019 Federated Conference on Computer Science and Information Systems (FedCSIS). pp. 61–65. IEEE (2019)
20. Qu, C., Zhang, T., Qiao, H., Tang, Y., Yuille, A.L., Zhou, Z., et al.: Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. Advances in Neural Information Processing Systems **36** (2024)

21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
22. Schmidt, A., Morales-Alvarez, P., Molina, R.: Probabilistic modeling of inter-and intra-observer variability in medical image segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21097–21106 (2023)
23. Shi, J., Zhang, K., Guo, C., Yang, Y., Xu, Y., Wu, J.: A survey of label-noise deep learning for medical image analysis. Medical image analysis **95**, 103166 (2024)
24. Strijbis, V.I., Gurney-Champion, O.J., Slotman, B.J., Verbakel, W.F.: Impact of annotation imperfections and auto-curation for deep learning-based organ-at-risk segmentation. Physics and Imaging in Radiation Oncology **32**, 100684 (2024)
25. Sudre, C.H., Anson, B.G., Ingala, S., Lane, C.D., Jimenez, D., Haider, L., Varsavsky, T., Tanno, R., Smith, L., Ourselin, S., et al.: Let's agree to disagree: Learning highly debatable multirater labelling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 665–673. Springer (2019)
26. Vorontsov, E., Kadoury, S.: Label noise in segmentation networks: mitigation must deal with bias. In: MICCAI Workshop on Deep Generative Models. pp. 251–258. Springer (2021)
27. 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)
28. Wesemeyer, T., Jauer, M.L., Deserno, T.M.: Annotation quality vs. quantity for deep-learned medical image segmentation. In: Medical Imaging (2021), https://api.semanticscholar.org/CorpusID:232271204
29. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. Medical image analysis **102**, 103547 (2025)
30. Yu, S., Chen, M., Zhang, E., Wu, J., Yu, H., Yang, Z., Ma, L., Gu, X., Lu, W.: Robustness study of noisy annotation in deep learning based medical image segmentation. Physics in Medicine & Biology **65**(17), 175007 (2020)