Modeling Multi-annotator Uncertainty as Multi-class Segmentation Problem

Martin Žukovec, Lara Dular(✉), and Žiga Špiclin

Faculty of Electrical Engineering, University of Ljubljana,
Tržaška cesta 25, 1000 Ljubljana, Slovenia
lara.dular@fe.uni-j.si

Abstract. Medical image segmentation is a monotonous, time-consuming, and costly task performed by highly skilled medical annotators. Despite adequate training, the intra- and inter-annotator variations results in significantly differing segmentations. If the variations arise from the uncertainty of the segmentation task, due to poor image contrast, lack of expert consensus, etc., then the algorithms for automatic segmentation should learn to capture the annotator (dis)agreements. In our approach we modeled the annotator (dis)agreement by aggregating the multi-annotator segmentations to reflect the uncertainty of the segmentation task and formulated the segmentation as multi-class pixel classification problem within an open source convolutional neural architecture nnU-Net. Validation was carried out for a wide range of imaging modalities and segmentation tasks as provided by the 2020 and 2021 QUBIQ (Quantification of Uncertainties in Biomedical Image Quantification) challenges. We achieved high quality segmentation results, despite a small set of training samples, and at time of this writing achieved an overall third and sixth best result on the respective QUBIQ 2020 and 2021 challenge leaderboards.

Keywords: Multi-class segmentation · Noisy labels · Uncertainty aggregation · Convolutional neural networks · Challenge datasets

1 Introduction

Image segmentation is one of the fundamental tasks of medical imaging, crucial in modeling normal patient anatomy, detection of pathology, analysing patient's health status and indicating medical treatments and procedures. For instance, manual segmentation prior to surgical tumor removal and organ-at-risk contouring for radiotherapy planning is a time-consuming, mundane and thus a costly task carried out by expert annotators.

Developing automated algorithms can greatly reduce both time and money spent on medical image segmentation tasks. However, it is of extreme importance to estimate the uncertainty of output segmentation, as poor segmentation may adversely impact upon based treatments and procedures. Despite extensive

A. Crimi and S. Bakas (Eds.): BrainLes 2021, LNCS 12962, pp. 112–123, 2022.
https://doi.org/10.1007/978-3-031-08999-2_9

expert training and experience, many researches found contours on common set of images to differ significantly between the experts [5]. These may naturally arise from the uncertainty of the segmentation task, due to poor image contrast, lack of expert consensus, etc. We therefore should expect the uncertainty of annotations to reflect in the predictions of automated algorithms.

The Quantification of Uncertainties in Biomedical Image Quantification (QUBIQ) challenge [10] aims to develop and evaluate automatic algorithms for quantification of uncertainties, arising from experts' (dis)agreement in biomedical image segmentation. In 2020 the challenge presented four different MR and CT image datasets on which a total of seven segmentation tasks were released. In 2021, the organisers added two datasets each with a single task.

This paper presents our approach to capturing multi-annotator segmentation uncertainty for nine tasks of the QUBIQ 2020 and 2021 challenges. First the multi-annotator segmentations were aggregated, considering the same performance level for each of the expert annotators, such that they approximate the segmentation task uncertainty. We advanced the state-of-the-art nnU-Net convolutional neural network (CNN) model by casting multi-annotator uncertainty estimation as multi-class segmentation problem, where aggregated segmentations were the prediction target. Thus the model was able to capture and recreate the experts' (dis)agreements. At the time of this writing[1] the proposed approach achieved the third and sixth best scores on the respective QUBIQ 2020 and 2021 leaderboards.

2 Related Work

Supervised machine learning models like the deep CNNs for image segmentation generally require large training datasets of annotated images to achieve adequate performance levels. In medical imaging domain, however, we typically obtain small datasets due to the high effort required to obtain expert annotations (i.e. manual segmentations). When training models with a single expert segmentation per image we typically consider it as ground truth (GT), despite potential annotator bias and noise. A natural strategy to reduce the impact of annotator bias and noise is to consider the annotations of multiple experts.

With the availability of multiple expert segmentations a common approach is to conceive a fusion strategy to approximate the GT [5]. The most straightforward approach is Consensus voting, annotating the area as GT if all annotators agree, and Majority voting [4,9], assigning pixel labels according to the majority rule. These definitions can be generalized by using different agreement levels. Lampert et al. [7] reported that increasing the level of agreement for forming GT increased the model's reported performance. They further noted that a higher agreement level could result in over-optimistic results, as this could be the consequence of choosing the most obvious segments of the region of interest (ROI). Further, the problem with such an approach is the loss of information about inter-annotator variability.

[1] September 9, 2021.

A more advanced and widely used approach to aggregating multiple expert segmentations is the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm proposed by Warfield et al. [13]. The STAPLE algorithm uses expectation-maximization to compute a probabilistic estimate of the true segmentation and the sensitivity and specificity performance characteristics for each annotator. A similar approach to STAPLE is used in SIMPLE [8] which additionally iteratively estimates the performance of segmentations and discards poorly performing segmentation before finally fusing the remaining segmentations to estimate GT. Lampert et al. [7] showed that STAPLE performs well when inter-annotator variability is low, but degrades with the increasing number of annotations and high variability of annotations. They also examined the effect of inter-annotator variance on foreground-background segmentation algorithms, in a computer vision setting. Despite not including deep neural networks, their results showed that the rank of the model is highly dependent on the chosen method used to form the GT. Furthermore, including a similar aggregation strategy into segmentation method will inevitably lead to overoptimistic results.

Training machine learning models on datasets with multiple segmentation masks in a supervised manner allows for different representations and uses of the input data for model training. Firstly, each image-segmentation pair can be treated as a separate sample. For instance Hu et al. [1] propose a segmentation model based on the Probabilistic U-Net [6], where during model training multi-annotator segmentations of each image were fed to the network in the same mini-batch. Zhang et al. [14] took into account the multi-annotator dataset in the construction of the model architecture. The so called U-Net-and-a-half was constructed from a single encoder and multiple decoders. Each decoder corresponded to an expert allowing for simultaneous learning from all masks. The loss function was computed as the aggregated loss across all decoders.

Many approaches model and/or quantify the segmentation output uncertainty. For instance, the model proposed by Hu et al. [1] based on the Probabilistic U-Net [6] uses inter-annotator variability as a training target. In this way, they were able to generate multiple diverse segmentations from each input, which represent a possible expert segmentation. Jungo et al. [4] computed uncertainty by the principle of Monte Carlo dropout. They used dropout layers at inference time to produce multiple segmentations and, by computing pixel-wise variance, estimated the model's uncertainty.

To summarize, when designing architectures for modeling annotator uncertainty on datasets with multiple annotations, we need to formulate several computational strategies: (i) a strategy to deal with multiple annotations per image in the model training input, (ii) a strategy to approximate the ground truth, and finally (iii) a strategy to model uncertainty on the model output.

In this paper, we focus on the first and third points, i.e. the strategy of handling multiple annotations per image and modeling of output uncertainty, while as for the second point we latently acknowledge that ground truth may not exist. Thus we propose to aggregate multiple annotations into a single mask and to treat each level of agreement as a separate class. Modeling multi-annotator

uncertainty as multi-class segmentation problem can be simply coupled with any multi class segmentation model. According to a recent review on noisy label handling strategies [5] and our literature review, to the best of our knowledge, such a simple but effective solution to annotation aggregation and uncertainty modeling has not yet been proposed.

3 Materials and Methods

3.1 Datasets

The QUBIQ 2020 challenge data consists of four 2D CT and MR datasets of different anatomies with seven segmentation tasks, where two of the datasets, namely *Prostate* and *Brain tumor* dataset, include multiple ROIs. The QUBIQ 2021 challenge is an extension, including two additional 3D datasets, *Pancreas* and *Pancreatic lesion*, where each patient went through two scans at two time points. Each of the images was segmented by multiple trained experts, with annotator count ranging from 2 to 7, depending on the particular dataset. Additional dataset information is given in Table 1 and a few examples are visualized in Fig. 1.

Table 1. Number of given samples in training and validation dataset.

Dataset	No. samples (Train/Val.)	No. structures	No. contours	No. modalities
Prostate	55 (48/7)	2	6	1
Brain growth	39 (34/5)	1	7	1
Brain tumor	32 (28/4)	3	3	4
Kidney	24 (20/4)	1	3	1
Pancreas	58 (40/18)	1	2	1
Pancreatic lesion	32 (22/10)	1	2	1

3.2 Multi-annotation Aggregation

For segmentation of ROI given multiple annotations, we aggregated the given binary segmentation masks into a single input mask M^{in} as

$$M^{in}(x, y) = \sum_{i=1}^{N} B_i(x, y), \tag{1}$$

where N denotes the number of experts and B_i the binary value of pixel (x, y) as annotated by i-th expert. The values of the encoded mask were thus between 0 and the number of experts, where each foreground pixel value denotes the number of experts labeling the selected pixel as the ROI. In this way we encode

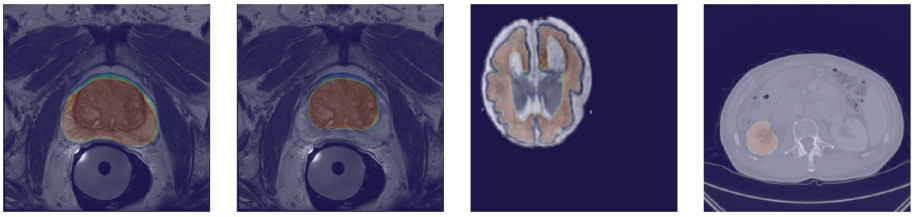

Fig. 1. Exemplary images with multi-annotator masks for segmentation tasks with single modalities (*left to right*): Prostate – Task 1, Prostate – Task 2, Brain growth and Kidney. The color notes the number of experts marking the area as segmented organ, from 0 (*blue*) to all (*red*). (Color figure online)

the three-dimensional mask input (no. of annotators, width, height) and map it into a two-dimensional space $[N\ X\ Y] \xrightarrow{M_{in}} [X\ Y]$, for image width X and image height Y, as shown in Fig. 2. By encoding multiple image masks, we transformed the problem into multi-class classification problem, with $N+1$ classes (including background), where class c marks the agreement of exactly c annotators, for $c \in \{0, 1, \ldots, N\}$.

Fig. 2. Encoding of three binary segmentation masks (a), (b) and (c) into a single encoded multi-annotator mask (d), where each pixel value equals to the number of experts marking the particular pixel as the ROI.

The CNN output is a three dimensional matrix $[N+1\ X\ Y]$, with a vector $(p_0, p_1, \ldots, p_N)$ for pixel (x, y), where p_i marks the probability that the pixel was marked as ROI by exactly c annotators. By computing $\text{argmax}_c p_c$ for each pixel, we get a two dimensional output mask M^{out} with predicted regions of agreement between experts. We can further decode the output mask into three-dimensional space, as shown in Fig. 3, we obtain as many masks as there are annotators, however, this time each mask represents a quantitative agreement value. Thus, the output mask M_1^{out} represents the area where the structure of interest has been annotated by at least one annotator, whereas the marked area decreases with increasing the index j in

$$M_c^{out}(x, y) = (M^{out}(x, y) \geq c). \tag{2}$$

The output mask M_1^{out} thus represents the pixels that would be marked by at least one annotator, M_2^{out} by at least two annotators, etc. Further, dividing the output mask values with the number of annotators N results in values on the interval $[0, 1]$, which can be interpreted as annotation or segmentation (un)certainty and reflects the uncertainty of the expert annotators.

Fig. 3. Decoding of model output a) M^{out} into three binary segmentation masks b) M_1^{out}, c) M_2^{out} and d) M_3^{out}, where M_c^{out} denotes the predicted mask with ROI marked by at least c experts.

3.3 Segmentation Model

Structure segmentation and its uncertainty estimation was obtained by adapting the open-source nnU-Net [2,3]. The nnU-net framework implements a single or cascaded U-net model and, based on the input images, the particular model and its hyperparameters are chosen and configured automatically. The following subsections describe the framework and its adaptations.

Model Architecture. The nnU-Net ('no-new-Net') uses a 2D U-net or 3D U-Net [11] as a backbone architecture. The main advantage of nnU-Net is it's self configuring training pipeline and automatic adaptation of model architecture and hyperparameter tuning that considers the available hardware resources and requires little or no user input. The encoder part starts with 32 feature maps in the initial layers and doubles the number of feature maps with each downsampling and vice versa in the decoder. The number of convolutional blocks is adapted to the input patch size, assuring that downsampling does not result in feature maps smaller than $4 \times 4 (\times 4)$. Compared to the original U-net, the nnU-net authors replaced the ReLU activation functions with leaky ReLU and batch normalization with instance normalization.

Loss Function. We applied the *soft Dice loss function* directly on CNN output probabilities. The output values were mapped to the $[0, 1]$ interval using the softmax activation function on the output layer. For each class $c \in \{0, 1, \ldots, N\}$, where class $c = 0$ represents the background without any annotations, we compute soft Dice similarity coefficient

$$sDSC_c = \frac{\sum_{x,y} p_c(x, y) \cdot M_c^{in}(x, y)}{\sum_{x,y} p_c(x, y)^2 + \sum_{x,y} M_c^{in}(x, y)^2}, \tag{3}$$

where $p_c(x, y)$ denotes the output probability of pixel (x, y) belonging to class c, M_c^{in} the binary input mask of class c. Finally, Dice coefficient was averaged over all $N + 1$ classes. For the loss function we take the negative value

$$Loss = -\frac{1}{N+1} \sum_{c=0}^{N} sDSC_c. \tag{4}$$

Model Training. For each of the nine segmentation tasks on six dataset we trained a separate nnU-Net model that converged on average in 50 epochs. A 2D model was trained for each of the 2D image segmentation task and a 3D model with full resolution for the two 3D segmentation tasks. Note that 2D models were trained also for 3D data, however, the 2D model performed worse than the 3D model. In the case of multi modal data, i.e. brain lesions, a single model was trained using all image modalities as the model input.

Based on the data fingerprint and a series of heuristic rules the image resampling and image normalization were determined. Further, the architecture of nnU-Net dynamically adapted to the dataset, selecting appropriate image input patch size and batch size [2]. To allow training on large image patches, the batch size was generally small, typically (but not less than) two images per batch.

The nnU-Net model training included various data augmentation transformations, each with certain probability p. Namely, random rotations ($p = 0.2$), scaling ($p = 0.2$), mirroring ($p = 0.2$), Gaussian noise ($p = 0.1$) and smoothing ($p = 0.2$), and additive or multiplicative inhomogeneity simulation ($p = 0.15$ or 0.15, respectively). Models were trained using stochastic gradient descent optimizer with an initial learning rate of 0.01 and Nesterov momentum of 0.99.

The nnU-Net models were trained on patches that overlapped by half of the patch size. During inference the same patch size was used as during training. The predicted patches were then combined such that the contributions of different patch predictions across the common voxels were aggregated by weighing the predictions based on the voxel location. Since accuracy was expected to drop towards the patch border, the contribution of such voxels was less then the pixels close to the patch center.

Finally, the predictions were postprocessed by first checking the training dataset samples if all classes lied within a single connected component. In this case, this property was also imposed to the test set by retaining the single largest connected component for each class.

3.4 Evaluation Metrics

Model performance was evaluated according to the provided evaluation code by the QUBIQ challenge organizers. We compared the predicted uncertainty mask M^{out}/N with the uncertainty of the GT, computed as M^{in}/N. For each image, the uncertainty masks were binarized at thresholds $0.1 \times i$; $i = 0, 1, \ldots, 9$, for which the Dice coefficient DSC_i was computed as

$$DSC_i = \frac{2TP_i}{2TP_i + FP_i + FN_i},$$

where TP denotes the true positive pixels, FP denotes the false positive pixels and FN the false negative pixels. Finally the scores were averaged across all ten values for the final performance estimation

$$\overline{DSC} = \frac{1}{10} \sum_{i=0}^{9} DSC_i.$$

4 Results

The results of our proposed model on six datasets and across nine segmentation tasks were computed on the validation datasets and are reported in Table 2 and Fig. 4. In four of the 2D segmentation tasks our approach achieved an average Dice score over 0.9, while for the other three 2D tasks it achieved a score of over 0.7. The lowest scores, significantly below the average, were achieved for the two 3D segmentation tasks introduced in QUBIQ 2021 challenge.

Table 2. Performance measure $\overline{DSC}$ per segmentation task evaluated by QUBIQ challenge organizers. The average is computed over seven tasks for QUBIQ 2020 (disregarding pancreas and pancreatic lesion) and over nine tasks for QUBIQ 2021. (Note: Evaluation metrics on QUBIQ 2020 and 2021 leaderboard are not identical. The average score of our model reported on QUBIQ 2020 leaderboard equals to 0.7476.)

Structure	$\overline{DSC}$
Brain growth	0.9336
Brain tumor - Task 1	0.9485
Brain tumor - Task 2	0.7808
Brain tumor - Task 3	0.7639
Kidney	0.9766
Prostate - Task 1	0.9610
Prostate - Task 2	0.8280
Pancreas	0.5605
Pancreatic lesion	0.3990
Average	
– QUBIQ 2020	0.8846
– QUBIQ 2021	0.7946

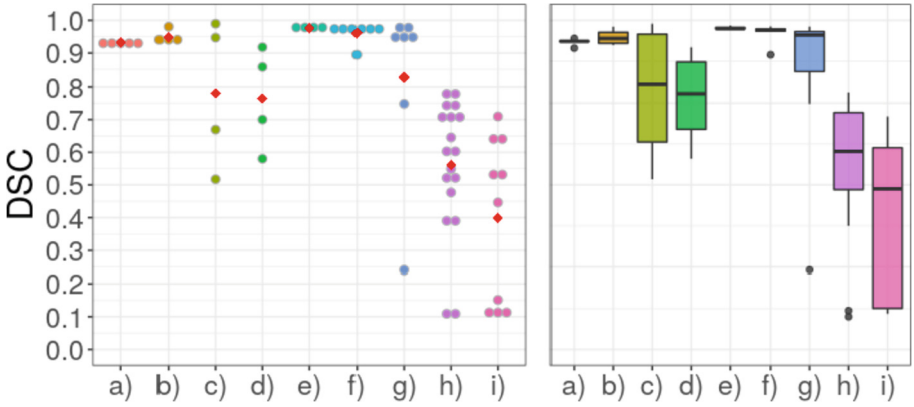

Fig. 4. Scatter plot (*left*) with marked mean values (*red*) and boxplot (*right*) of individual values of the average Dice coefficient $\overline{DSC}$ on validation images for seven tasks: *a)* Brain growth, *b)* Brain tumor - Task 1, *c)* Brain tumor - Task 2, *d)* Brain tumor - Task 3, *e)* Kidney, *f)* Prostate - Task 1, *g)* Prostate - Task 2, *h)* Pancreas, *i)* Pancreatic lesion. (Color figure online)

Due to the nature of the metric $\overline{DSC}$, the error of an incorrectly predicted pixel can accumulate when computing the DSC across multiple thresholds. In some validation images from the Brain tumor dataset (Tasks 2 and 3) the area of the annotated structures, where a fraction of the experts agree, measured only a few pixels. For such a small area an incorrect output value of even a single pixel changes the $\overline{DSC}$ metric value by a substantial amount.

Structures *a)* Brain growth, *b)* Brain tumor - Task 1, *e)* Kidney and *f)* Prostate - Task 1 were predicted consistently, without significant variation in the $\overline{DSC}$ between different cases. Due to the consistent labelling of all experts and consistent size of the structures, the neural network predictions were also consistent. In the case of the listed structures, the region of agreement of all experts was much larger than the region of disagreement, compared to the other structures. In practice, this means that the misperceived agreement pattern of a subset of annotators does not contribute to the value of metric $\overline{DSC}$ to the extent that it does in the case of small structures.

For the two 3D segmentation tasks, i.e. *h)* Pancreas and *i)* Pancreatic lesion, we observed a large variation of the $\overline{DSC}$ values. Specifically, in the cases with the value of $\overline{SDC}$ equal to 0.1, the model did not segment the ROI and instead returned an empty mask.

5 Discussion and Future Work

Intra- and inter-annotator variations result in significantly differing manual segmentations, which may be related to the uncertainty of the segmentation task; hence, the algorithms for automatic segmentation should learn to capture the annotator (dis)agreements. In our approach we modeled the annotator

(dis)agreement by aggregating the multi-annotator segmentations to reflect the uncertainty of the segmentation task and formulated the segmentation as multi-class pixel classification problem within an open source nnU-Net framework [3].

Validation was carried out for a wide range of imaging modalities and segmentation tasks as provided by the 2020 and 2021 QUBIQ challenges and showed high quality segmentations according to the average Dice scores. While inspecting our results we noticed a large variation in Dice scores across validation cases for the 2D Brain tumor segmentation tasks 2 and 3 and both segmentation tasks on the 3D datasets. In part, the low $\overline{DSC}$ scores in particular cases and high variability in the score in the aforementioned tasks can be attributed to the fact that the area of agreement covers only a few pixels. This is particularly evident for Brain tumor segmentation - Task 2, as shown in Fig. 5, where one of the raters consistently segments different ROIs as the other two raters. This systematic difference is also captured by the model, that did not classify any of the pixels as the area, where all the three annotators would agree.

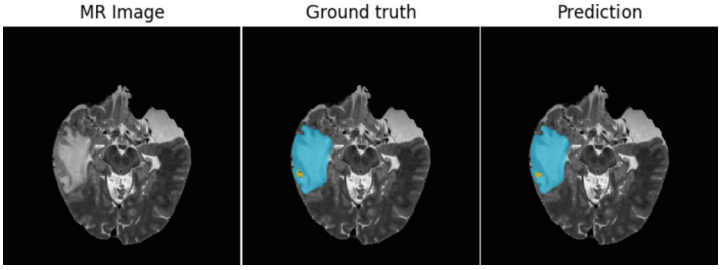

Fig. 5. Ground truth and prediction for Brain tumor - Task 2. Single annotator's ROIs marking *(blue)* significantly differ from the segmentation of the other two *(yellow)*, with a very small overlap of all three *(red)*. (Color figure online)

In 3D space, the ratio between the background and ROI becomes even larger. The poor result can therefore again be partially contributed to class imbalance. Further we can notice a large difference in input image sizes in z axis, that varies from 36 to 194 pixels on the training set. To potentially improve the results, reducing the image size around the ROI, before training the neural network could be considered.

When forming the aggregated target segmentation, we assumed, that all experts were equally trained and thus we took the sum of their segmentation masks as the ground truth. However, in the case of major disagreements in annotations, such as for Brain tumor segmentation - Task 2, a smaller weight could be given to the annotator that is not in accordance with the others. The performance might therefore be improved by the use of expert performance level estimates as obtained from the SIMPLE algorithm [8], confusion matrices as in Tanno et al. [12] or similar approaches for generating GT used as target masks.

Finally, one of the main limitations of modeling multi-annotator (dis)agreement as multi-class problem is it's sensitivity to minor changes of

the softmax function, which can result in pixel misclassification. A change of argmax function by, for example, weighted sum of classes using softmax outputs as weights, could result in a more robust model.

6 Conclusion

The goal of the QUBIQ challenge was to segment nine different structures of interest, i.e. organs and pathologies, in six different datasets, for which segmentation masks of multiple experts were provided. In the context of the established nnU-Net segmentation framework, we proposed a novel strategy of handling multiple annotations per image and modeling of output uncertainty. Namely, we aggregate multiple annotations into a single mask and to treat each level of agreement as a separate class, thus modeling multi-annotator segmentation uncertainty as multi-class segmentation problem. We achieved high quality segmentation results with an overall third and sixth best overall Dice score result on the respective QUBIQ 2020 and 2021 challenge leaderboards.

References

1. Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., Welling, M.: Supervised uncertainty quantification for segmentation with multiple annotations. arXiv:1907.01949 [cs, stat], July 2019. http://arxiv.org/abs/1907.01949
2. Isensee, F., Jäger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: Automated design of deep learning methods for biomedical image segmentation. Nat Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z, http://arxiv.org/abs/1904.08128, arXiv: 1904.08128 version: 2
3. Isensee, F., et al.: nnU-Net: self-adapting framework for U-net-based medical image segmentation. arXiv:1809.10486 [cs], September 2018. http://arxiv.org/abs/1809.10486
4. Jungo, A., Meier, R., Ermis, E., Herrmann, E., Reyes, M.: Uncertainty-driven sanity check: application to postoperative brain tumor cavity segmentation. arXiv:1806.03106 [cs], June 2018. http://arxiv.org/abs/1806.03106
5. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med. Image Anal. **65**, 101759 (2020). https://doi.org/10.1016/j.media.2020.101759
6. Kohl, S.A.A., et al.: A probabilistic U-net for segmentation of ambiguous images. arXiv:1806.05034 [cs, stat], January 2019. http://arxiv.org/abs/1806.05034
7. Lampert, T.A., Stumpf, A., Gançarski, P.: An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. IEEE Trans. Image Process. **25**(6), 2557–2572 (2016). https://doi.org/10.1109/TIP.2016.2544703, http://arxiv.org/abs/1307.0426
8. Langerak, T.R., van der Heide, U.A., Kotte, A.N.T.J., Viergever, M.A., van Vulpen, M., Pluim, J.P.W.: Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Trans. Med. Imaging **29**(12), 2000–2008 (2010). https://doi.org/10.1109/TMI.2010.2057442

9. Litjens, G., Debats, O., van de Ven, W., Karssemeijer, N., Huisman, H.: A pattern recognition approach to zonal segmentation of the prostate on MRI. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012. LNCS, vol. 7511, pp. 413–420. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33418-4_51
10. Quantification of Uncertainties in Biomedical Image Quantification Challenge 2021. https://qubiq21.grand-challenge.org/. Accessed 11 Aug 2021
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. arXiv:1505.04597 [cs], May 2015. http://arxiv.org/abs/1505.04597
12. Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D.C., Silberman, N.: Learning from noisy labels by regularized estimation of annotator confusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019
13. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging **23**(7), 903–921 (2004). https://doi.org/10.1109/TMI.2004.828354, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1283110/
14. Zhang, Y., et al.: U-net-and-a-half: convolutional network for biomedical image segmentation using multiple expert-driven annotations. arXiv:2108.04658 [cs], August 2021. http://arxiv.org/abs/2108.04658