An Exploration of Uncertainty Information for Segmentation Quality Assessment

Katharina Hoebel[a,b], Vincent Andrearczyk[c], Andrew Beers[a], Jay Patel[a,b], Ken Chang[a,b], Adrien Depeursinge[c,d], Henning Müller[c,e], and Jayashree Kalpathy-Cramer[a]

[a]Athinoula A. Martinos Center for Biomedical Imaging, Charlestown, MA, USA
[b]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA
[c]University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
[d]Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland
[e]University of Geneva, Switzerland

ABSTRACT

Including uncertainty information in the assessment of a segmentation of pathologic structures on medical images, offers the potential to increase trust into deep learning algorithms for the analysis of medical imaging. Here, we examine options to extract uncertainty information from deep learning segmentation models and the influence of the choice of cost functions on these uncertainty measures. To this end we train conventional UNets without dropout, deep UNet ensembles, and Monte-Carlo (MC) dropout UNets to segment lung nodules on low dose CT using either soft Dice or weighted categorical cross-entropy (wcc) as loss functions. We extract voxel-wise uncertainty information from UNet models based on softmax maximum probability and from deep ensembles and MC dropout UNets using mean voxel-wise entropy. Upon visual assessment, areas of high uncertainty are localized in the periphery of segmentations and are in good agreement with incorrectly labelled voxels. Furthermore, we evaluate how well uncertainty measures correlate with segmentation quality (Dice score). Mean uncertainty over the segmented region (U_{labelled}) derived from conventional UNet models does not show a strong quantitative relationship with the Dice score (Spearman correlation coefficient of -0.45 for the soft Dice vs -0.64 for the wcc model respectively). By comparison, image-level uncertainty measures derived from soft Dice as well as wcc MC UNet and deep UNet ensemble models correlate well with the Dice score. In conclusion, using uncertainty information offers ways to assess segmentation quality fully automatically without access to ground truth. Models trained using weighted categorical cross-entropy offer more meaningful uncertainty information on a voxel-level.

Keywords: computer vision, lung nodule, segmentation, uncertainty

1. INTRODUCTION

As part of developing trust in artificial intelligence models and to pave the way for their eventual adoption into the clinical workflow, it is important to understand how a model comes to its decisions and how reliable these are. To date, most deep learning methods for the automatic segmentation of target structures from medical imaging treat the problem as a supervised mapping between the input image and a binary output. However, due to the absence of clearly defined boundaries of targets such as tumors, manual outlines used as the ground truth suffer from high inter- and intra-rater variability[1,2] As a consequence, training data for algorithm development resembles weak labels rather than an absolute ground truth, increasing the model's uncertainty. If a segmentation algorithm in the clinic or in a research setting provided information about its confidence in its output, trust in the model and the downstream analysis would be greatly improved.

A model's uncertainty can be divided into aleatoric and epistemic uncertainties. While epistemic uncertainty, the uncertainty over the true values of a model's weights, decreases with increasing number of training data,

Further author information: (Send correspondence to K. Hoebel)
E-mail: khoebel@mit.edu

aleatoric uncertainty represents an intrinsic measure of the noise in the data, e.g. inter-rater variability, and is unaffected by the amount of available data.[3]

Recent work by Hu et al., has leveraged datasets with multiple annotations per training sample to model aleatoric and epistemic uncertainty separately.[4] However, even in the absence of multiple annotations, we are still able to extract the overall uncertainty of a model in its predictions by using uncertainty maps provided as a natural measure of uncertainty by Bayesian deep networks. In multi-stage models these uncertainty estimates can be propagated and used in downstream analysis steps to improve the overall performance of a pipeline e.g. for segmentation or pulmonary nodule detection.[5,6] Furthermore, recent work by Roy et al. and DeVries et al. has shown that image-level uncertainty estimates derived from a model's uncertainty maps correlate with the quality of a segmentation.[7,8]

Here, we explore the use of uncertainty measures on 3D segmentation of lung nodules. We compare three methods for uncertainty estimation: Maximum softmax probability derived from the output of a UNet,[9] and two methods that approximate Bayesian neural networks: deep ensembling[10] and Monte-Carlo (MC) dropout UNet.[11,12] Furthermore, we assess how the choice of cost function influences the uncertainty estimates derived from these models.

2. METHODS

2.1 Data set

The data set used for this study consists of a subset of the American National Lung Screening Trial (NLST) of low dose CT scans of the thorax.[13] The lung nodules were manually segmented by radiologists at the University Hospitals of Geneva (HUG).[14] A total of 484 (244 benign/240 malignant, determined on histology) of 450 individual patients' nodules were annotated (one nodule per study). The data set is split on the patient level into training, validation and test data sets containing 316 (166 benign/150 malignant), 67 (30/37), and 101 (48/53) cases respectively.

Table 1. data set composition

Data set	N patients	N cases (benign/malignant)
Training	316	316 (166/150)
Validation	67	67 (30/37)
Test	67	101 (48/53)

2.2 Lung Nodule Segmentation

The cropped volumes are windowed between -1000 and 400 HU and each volume is standardized (zero mean, unit variance). We train 3D segmentation models on patches of size 32x32x16 and additional data augmentation is implemented through sagittal flipping.

We explore the uncertainty information derived from three different models: maximum softmax probability from standard UNet models,[9,15] deep UNet ensembles,[10] and MC dropout UNets.[11,12] Deep ensembles consist of five separate UNet semgentation models, all randomly initialized and trained on the full training dataset. Each network is intended to produce slightly varying segmentations. These predictions can successively be treated as independent samples similar to samples derived from Bayesian networks. The third variation of segmentation models are Monte Carlo dropout UNets. By applying dropout after each convolutional layer during training and test time, the MC dropout UNet approximates probabilistic neuron connectivity similar to a Bayesian neural network.

As cost functions, we use either soft Dice

$$d(p,g) = 1 - D(b(p),g) = 1 - \frac{2\sum_x (b(p(x)\cdot g(x)))}{\sum_x b(p(x)) + \sum_x g(x) + 1},$$

or weighted categorical cross-entropy

$$wcc(p,g) = \sum_x \left(-\alpha g(x) \log p(x) + \beta(1-g(x)) \log(1-p(x))\right),$$

where $p(x)$ denotes the model's output for each pixel to belong to a lung nodule x, $g(x)$ the ground truth, $D(p,g)$ the Dice score between p and g, b the binarization function with threshold 0.5, and $\alpha = 3.0$, $\beta = 0.1$ are the weighting factors to up-weight the foreground class (α) and down-weight the background class (β).

The deep learning models were implemented in DeepNeuro[16] (Keras with Tensorflow backend) and trained on an NVIDIA K80 graphics processing unit.

We sample $N_{ensemble}$=5 times from deep ensembles consisting of 5 models, and N_{MC}=20 from MC dropout UNets. To generate a final segmentation result for deep ensembles and MC dropout models, we generate N slightly different samples for each input, average the voxelwise probability over these samples to generate a final segmentation probability map and then binarize this map.

2.3 Quantification of Uncertainty Information

For the conventional UNet models, we derive voxel-wise segmentation uncertainty $U_{UNet}(x)$ based on the distance between the output of the softmax activation function and the binarization threshold $t = 0.5$:

$$U_{\text{UNet}}(x) = 1 - |1 - 2p(x)|$$

(referred to as maximum softmax uncertainty). For deep ensembles and MC dropout models, the voxel-wise segmentation uncertainty is estimated as the mean entropy over all N samples:

$$U_{\text{ensemble, MC}}(x) = -\frac{1}{N} \sum_{i=1}^{N} p_i(x) \log(p_i(x)).$$

Image-level uncertainty is described by mean uncertainty over a segmentation:

$$U_{\text{labelled}} = \frac{1}{\sum_x b(p(x))} \sum_{p(x) \geq 0.5} U(x)$$

for the output of conventional UNet, deep UNet ensemble, and MC dropout UNet models. Additionally, we derive two more uncertainty measures for the output of deep UNet ensembles and MC dropout UNet models. The following uncertainty measures are computed based on the N samples drawn for each input as described by Roy et al.[7] coefficient of variation:

$$CV = \frac{\text{Var}_i \left(\sum_x b(p_i(x))\right)}{\text{E}_i \left[\sum_x b(p_i(x))\right] + 1},$$

and mean pairwise Dice:

$$D_{pw} = \text{E}\left[\{D(b(p_i(x)), b(p_j(x)))\}_{j>i}\right].$$

2.4 Segmentation quality prediction

A linear regression model is developed to predict segmentation quality by means of Dice score between segmentation and ground truth based on the three ground truth independent uncertainty measures, CV, D_{pw}, and $U_{labelled}$, using the python scikit-learn package.

3. RESULTS

3.1 Lung Nodule Segmentation Models

We developed a total of six lung nodule segmentation models using a combination of three different techniques: UNet, deep UNet ensembles, and MC dropout UNets and two different cost functions: soft Dice and wcc. The best performance for MC dropout UNets was achieved using a dropout rate of $p_{MC} = 0.2$. The final segmentation prediction for deep ensembles and MC dropout models was generated as described in Section 2.2. The average Dice scores for the predictions on the test data set range from 0.70 (Dice UNet) to 0.77 (wcc deep UNet ensemble) and are listed in Table 2. For all three methods, the models trained using wcc perform better than the ones trained with soft Dice. Rows 1, 3, and 5 in Figure 1 illustrate the segmentation results of some exemplary cases.

Table 2. Segmentation performance of the six developed lung nodule segmentation models (conventional UNet, deep ensemble, and MC dropout UNet each trained using soft Dice or wcc as cost function) on the independent test dataset.

Model	soft Dice	wcc
UNet	0.70	0.71
Deep UNet ensemble	0.74	0.77
MC dropout UNet	0.70	0.74

3.2 Qualitative Assessment of the Segmentation Uncertainty Distribution

Voxelwise uncertainty maps for all models are generated as described in 2.3. On visual assessment, the uncertainty maps based on the maximum softmax probability of both UNet models as well as the uncertainty maps for deep UNet ensembles and MC dropout UNets show high uncertainties in the periphery of the segmentations (Figure 1, row 2, 4, and 6 with their respective segmentation results above). Across all three methods, the uncertainty maps derived from models trained using soft Dice have a steep uncertainty gradient at the margins, whereas the wcc models' uncertainty maps have a larger area of high uncertainty in the periphery of the segmentations with a less steep gradient between areas of high and low uncertainty. Especially in the case depicted in the first and second row of Figure 1, areas of high uncertainty correspond to areas that are false positive or false negative (yellow and red labels in row 1). This effect seems to be even more pronounced for models trained using wcc (columns 3, 5, and 7). However, in many cases, false positive areas are associated with low uncertainties, comparable with these of true positive areas. This effect appears more distinct for models trained with soft Dice. Comparing the different methods visually, uncertainty maps derived from MC dropout models appear to be less noisy than the ones from UNets and deep UNet ensembles (especially for models trained using wcc).

3.3 Quantitative Assessment of Segmentation Uncertainty

To obtain more quantitative information about the relationship between the quality of a predicted segmentation and the uncertainty distribution, we assessed the correlation between the Dice score of a segmentation (as image quality metric) and different image level uncertainty measures computed from voxelwise uncertainty maps or variation between samples (for deep ensembles and MC dropout) as a surrogate for predictive uncertainty.

UNet The only ground truth independent uncertainty measure derivable from the maximum softmax uncertainty maps is the mean uncertainty over all voxels of the predicted segmentation ($p(x) \geq .5$), U_{labelled}. On the test data set, however, U_{labelled} did not show good correlation with the Dice score between ground truth and the binarized segmentation map (Spearman correlation coefficient -0.45/-0.64 for the Dice and wcc UNet respectively).

In addition to the ground truth independent mean uncertainty over the segmented region, we examined whether the mean uncertainty over true positive ($U_{\text{true pos}}$) areas differs from false positive/false negative areas ($U_{\text{false pos}}$, $U_{\text{false neg}}$) (green/yellow/red labels in Figure 1). Indeed, for the models trained using soft Dice, $U_{\text{true pos}}$ shows a stronger relationship with the segmentation Dice (-0.46) compared to $U_{\text{false pos}}$ (-0.27) and $U_{\text{false neg}}$ (0.16). In contrast, the maximum softmax uncertainty derived from the wcc UNet model shows overall a stronger correlation between the uncertainty over the labelled region. Here, this relationship is driven by a high correlation coefficient with true positive labelled regions (-0.74), while there is no evident connection between segmentation quality and $U_{\text{false pos}}$ and a moderately strong correlation with $U_{\text{false neg}}$.

Deep UNet ensemble and Monte Carlo dropout UNet For the developed deep UNet ensembles and MC dropout UNet models, we examined the usability of three image-level uncertainty measures: coefficient of variation (CV), mean pairwise Dice score between the N predictions (D_{pw}),and mean uncertainty of the segmentation (U_{labelled}) described in 2.3. The correlation coefficients for all uncertainty measures and models are listed in Table 4. Comparing the strength of the relationship between segmentation quality and the different

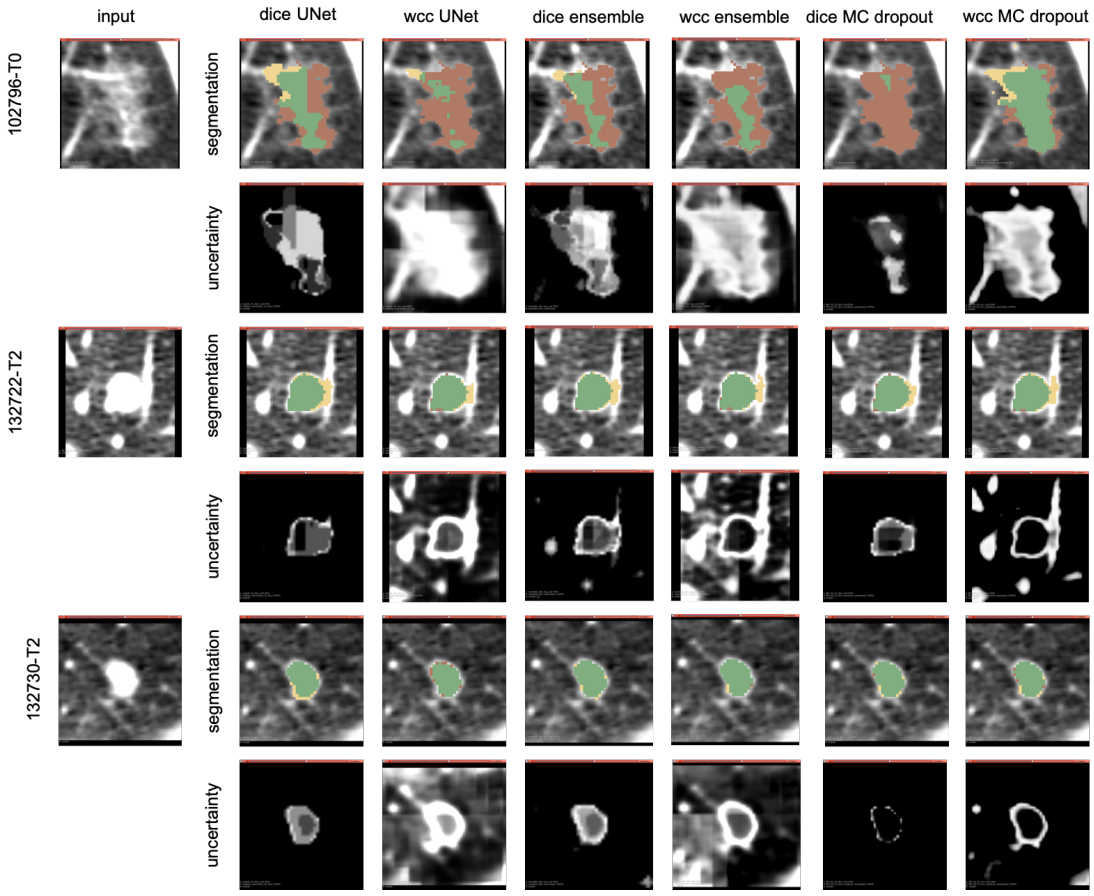

Figure 1. Segmentation predictions (green: true positive, yellow: false positive, red: false negative) and spatial uncertainty maps (brighter areas correspond to higher uncertainty) for the six developed lung nodule segmentation models.

Table 3. Spearman correlation coefficient between segmentation Dice score and measures of uncertainty for the UNet models on the test data set, p values in parentheses.

Model	all labelled	true positive	false positive	false negative
Dice UNet	-0.46 (5.6e-06)	-0.46 (4.5e-06)	-0.27 (9.9e-03)	0.16 (0.13)
wcc UNet	-0.64 (9.2e-12)	-0.74 (1.1e-16)	0.01 (0.95)	0.56 (9.7e-09)

uncertainty measures extracted from deep UNet ensembles and MC dropout UNets (separately for Dice and wcc), the correlations with uncertainty measures derived from MC dropout models are stronger than for deep ensembles. This effect is also present in the same magnitude between the coefficients from our deep ensembles with uncertainty measures computed from only five MC dropout samples (data not shown). While CV and D_{pw} show comparable (if slightly higher for Dice) correlations with the segmentation quality for all deep ensemble and MC dropout models, there is a clear advantage for models trained using wcc for U_{labelled} (-0.43/-0.49 vs -0.74/-0.82 for Dice vs wcc deep ensemble/MC dropout models).

Dependence of Uncertainty Metrics on the Target Size As described in the previous section, the areas of highest uncertainty are concentrated around the margins of the predicted segmentation. Accordingly, we examined whether there is a confounding dependence between the described three uncertainty metrics and the surface area or volume of a segmentation. The Spearman correlation coefficients between uncertainty metrics

Table 4. Spearman correlation coefficient between segmentation Dice score and measures of uncertainty, p values in parentheses

Model	CV	D_{pw}	U_{labelled}
Dice deep UNet ensemble	-0.61 (2.0e-11)	0.70 (1.7e-15)	-0.43 (9.4e-06)
wcc deep UNet ensemble	-0.68 (1.2e-14))	0.74 (1.4e-18)	-0.74 (3.7e-18)
Dice MC dropout UNet	**-0.78** (1.56e-21)	**0.83** (4.4e-27)	-0.49 (1.48e-07)
wcc MC dropout UNet	-0.75 (1.3e-19)	0.79 (1.2e-22)	**-0.82** (1.0e-25)

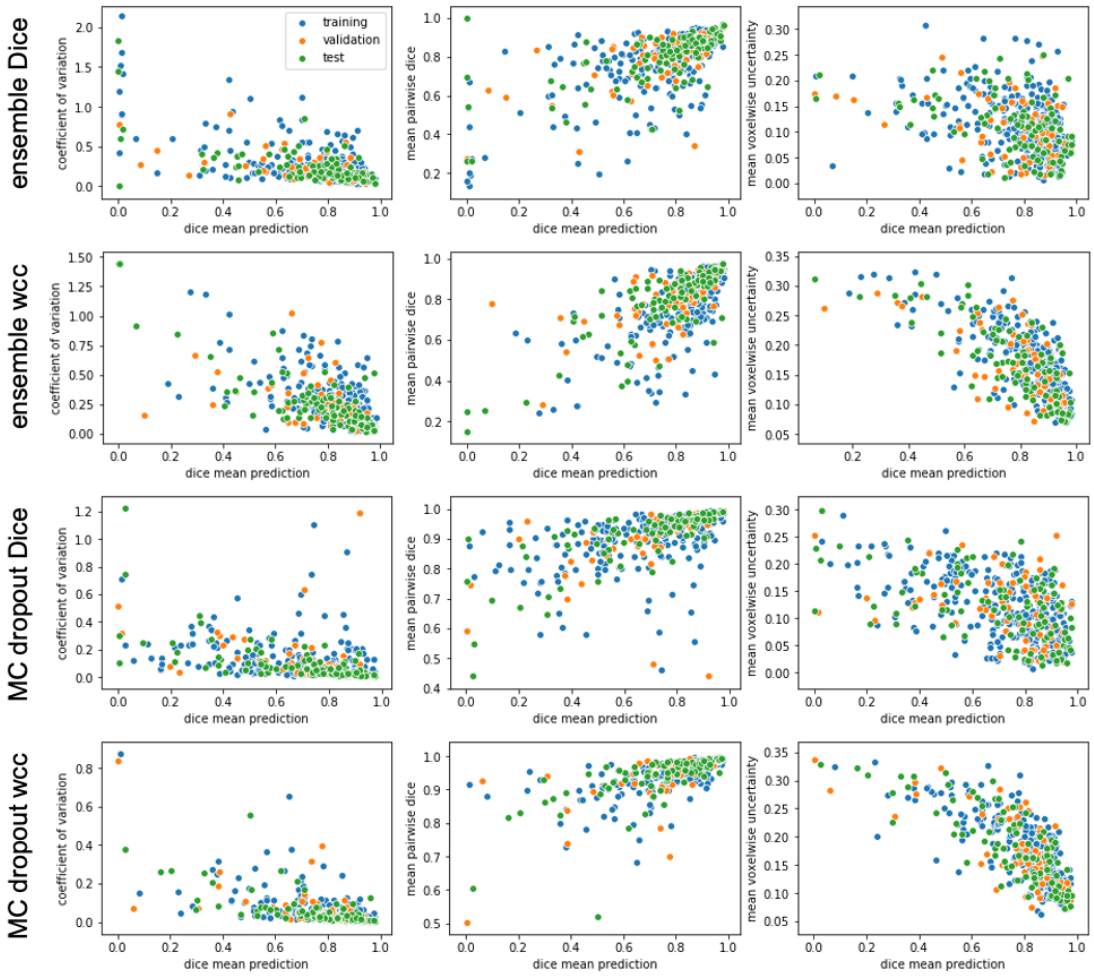

Figure 2. Correlation between segmentation quality (Dice score) and uncertainty measures for deep UNet ensembles (row 1 and 2) and MC dropout UNet (row 3 and 4) models trained using (soft) Dice or wcc as cost function (training, validation, and test data set)

derived from the wcc MC dropout UNet and segmentation surface area/volume are listed in Table 5. None of the uncertainty metrics shows a significantly high correlation with either surface area or volume. Therefore, we restrain to correct uncertainty measures for surface area or volume.

Table 5. Spearman correlation coefficient between segmentation size measures (in units of voxels) and measures of uncertainty, p values in parentheses

Size Measure	CV	D_{pw}	U_{labelled}
surface area	0.01 (0.89)	-0.21 (3.0e-02)	0.25 (1.2e-02)
volume	-0.09 (0.39)	-0.12 (0.26)	0.15 (0.14)

3.4 Predicting Segmentation Quality in the Absence of Ground Truth

The ability to predict the segmentation quality independently of a ground truth would allow to flag a model's predictions for human review if they do not fulfill a set quality criterion. We develop a linear regression model based on uncertainty metrics extracted from the wcc MC dropout UNet to predict the Dice score of a predicted segmentation with the goal to use this model to flag cases with a predicted Dice score < 0.8 for human review. The model was developed on the training data set and evaluated on the independent test set. This preliminary model has a sensitivity of 87%, specificity of 60% and a false negative rate of 13%. Figure 3 shows the relationship between the true Dice score and the predicted Dice score for the cases in the test data set.

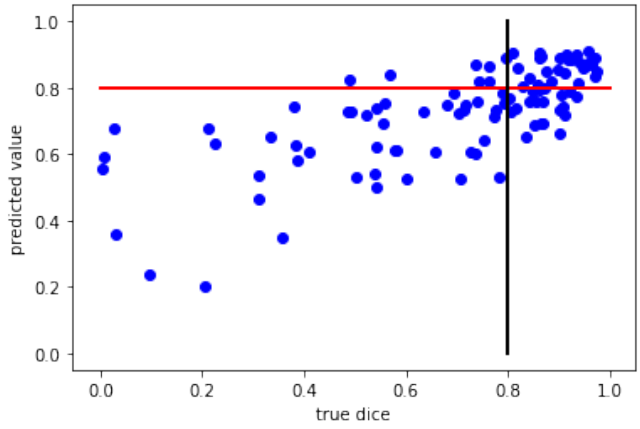

Figure 3. Linear regression model to predict the Dice score of a segmentation prediction (here test data set) from its uncertainty measures. The red line indicates the chosen cutoff for flagging of 0.8 and the red line the true decision boundary of 0.8.

4. DISCUSSION

We evaluated three different deep learning segmentation methods in combination with estimation of segmentation uncertainty for the segmentation of lung nodules from low dose CT imaging. Additionally, we compared the effect of two different widely used cost functions, soft Dice and weighted categorical cross-entropy (wcc), on the uncertainty measures extracted from the models' output. All six segmentation models perform well on the segmentation task (see Table 2) with models trained using wcc performing slightly better in comparison to models trained using soft Dice. This effect that is more accentuated for deep ensembles and MC dropout models than the conventional 3D UNets.

In contrast to some recently proposed methods,[4,17] the methods used for this work are not intended to maximize the variability in segmentation predictions by producing segmentation hypotheses that reflect the full space of potential segmentations rather than provide information about the models' uncertainty in addition to one most likely segmentation. Furthermore, they only require one set of annotations for the training data.

Qualitative Assessment of Uncertainty Maps The spatial uncertainty distributions derived from all six lung nodule segmentation models showed good agreement with areas of false positive and false negative segmentation. Generally, the areas of highest uncertainties are concentrated in the periphery of a segmentation. This is in good agreement with findings previously reported by other groups.[6–8,18] Qualitatively, the uncertainty maps derived from the wcc models of all three methods appear to capture the uncertainty distribution better than Dice models. Furthermore, the uncertainty maps are less noisy for MC dropout models as compared to conventional UNets and deep ensembles. While the qualitative visual assessment offers the potential to give researchers some insight into why and how a segmentation model tends to fail in certain cases, in the absence of a ground truth expert review of every single case is required. Therefore, we assessed how well ground truth independent uncertainty measures derived from a model's predictions correlate with the quality of the segmentation reflected by its Dice score.

Correlation Between Segmentation Quality and Ground-Truth Independent Uncertainty Measures While the mean uncertainty U_{labelled} derived from the maximum softmax probability of conventional UNets shows a moderate relationship with the Dice score of the prediction, this metric might be sufficient to give an idea in extreme cases (e.g. very good or very bad segmentation quality). However, this correlation is hardly sufficient to inform downstream tasks e.g. flagging cases for human review especially for borderline cases close to the chosen threshold.

In contrast, ground truth independent uncertainty measures computed from samples derived from deep UNet ensembles and MC dropout UNets show a strong relationship with segmentation quality. Both coefficient of variation and the mean pairwise Dice correlate well with the Dice score for models trained using Dice as well as wcc loss. However, this is not the case for the mean uncertainty over the labelled region: here, models trained using wcc reveal a much stronger correlation with segmentation quality as compared to models trained using Dice (for both deep ensembles as well as MC dropout UNets). This might be due to the smaller region of high uncertainty in Dice models as compared to wcc models (compared in Figure 1). The use of Dice as cost function seems to enforce steep probability gradients at the borders of a segmentation in contrast to a voxel-level cost function as wcc. Accordingly, for uncertainty measures based on the voxels' uncertainty values like the mean uncertainty over the segmentation, the use of a voxel-level cost function like wcc is advantageous over soft Dice resulting in both higher quality segmentation and image level uncertainty information.

In addition to providing uncertainty measures that show a higher correlation to segmentation Dice (Table 3.3), training and using MC dropout is slightly more convenient than deep ensembles as we are only required to train one model that allows to sample from it as often as desired. Our results for the correlation between segmentation quality and uncertainty measures are in good agreement with the results reported by Roy et al. for the segmentation of brain structures from MRI using an MC dropout segmentation model.[7] The model that provides the best overall uncertainty measures is the wcc MC dropout UNet.

Even though the concentration of the areas of highest uncertainty around the margins of the predicted segmentations (as illustrated by the uncertainty maps in row 2, 4, and 6 in Figure 1) might imply some relationship between the volume or surface area of a segmentation, our results imply that there is no need to correct uncertainty measures for the surface area or volume (Table 5). However, the correlation coefficient of -0.21/0.25 between the surface area of a predicted segmentation and mean pairwise Dice/mean uncertainty indicate that there is a weak connection and in other use cases a correction might be appropriate.

Predicting Segmentation Quality from Uncertainty Metrics Independent of a Ground Truth One potential application of uncertainty measures in deep learning segmentation pipelines would be the automatic flagging of segmentation predictions with an expected low segmentation quality for human review independent of the availability of a ground truth. Due to the continuous nature of the Dice score as segmentation quality metric (and other related metrics such as Hausdorff distance), a linear regression model is the preferred method over classification models such as support vector machine or random forest. As illustrated in Figure 3, the predicted segmentation Dice score is in good accordance with the true Dice score for cases in the independent test data set. Here, we are mostly concerned about the false negative cases in the left upper quadrant of Figure 3 - cases

that actually have a Dice score that is lower than the chosen threshold but where the predicted score is above the threshold. However, the true Dice of 5 out of the 7 false negative cases in our test data set is above 0.7, close to the chosen threshold of 0.8. Therefore, a misclassification, although not desirable, would not have a large impact on the overall performance in the setting of a large scale study. An additional advantage of a linear regression model over classification models is that the cutoff for cases that are routed for human review can be chosen arbitrarily without the need to re-train the regression model. For sure, if the result of a segmentation had an impact on e.g. the therapy a patient is about to receive, an automated quality assessment could not replace the careful review of an expert.

Future research should address the possibility to extract more uncertainty information from the spatial distribution of uncertainty over a segmentation and explore the use of the above described uncertainty measures for other disease sites and imaging modalities.

ACKNOWLEDGMENTS

This publication was supported from the Martinos Scholars fund to K. Hoebel. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Martinos Scholars fund.

This project was supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under award number 5T32EB1680 to K. Chang and J. Patel and by the National Cancer Institute of the National Institutes of Health under Award Number F30CA239407 to K. Chang. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

This study was supported by National Institutes of Health grants U01-CA154601, U24-CA180927, and U24-CA180918 to J. Kalpathy-Cramer.

This research was carried out in whole or in part at the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital, using resources provided by the Center for Functional Neuroimaging Technologies, P41EB015896, a P41 Biotechnology Resource Grant supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institutes of Health.

REFERENCES

[1] Chisholm, R. A., Stenning, S., and Hawkins, T. D., "The accuracy of volumetric measurement of high-grade gliomas," *Clinical Radiology* **40**(1), 17–21 (1989).

[2] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., Corso, J. J., Criminisi, A., Das, T., Delingette, H., Demiralp, C., Durst, C. R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K. M., Jena, R., John, N. M., Konukoglu, E., Lashkari, D., Mariz, J. A., Meier, R., Pereira, S., Precup, D., Price, S. J., Raviv, T. R., Reza, S. M. S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.-C., Shotton, J., Silva, C. A., Sousa, N., Subbanna, N. K., Szekely, G., Taylor, T. J., Thomas, O. M., Tustison, N. J., Unal, G., Vasseur, F., Wintermark, M., Ye, D. H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., and Van Leemput, K., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS).," *IEEE transactions on medical imaging* **34**, 1993–2024 (10 2015).

[3] Kiureghian, A. D. and Ditlevsen, O., "Aleatory or epistemic? Does it matter?," *Structural Safety* **31**, 105–112 (3 2009).

[4] Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., and Welling, M., "Supervised Uncertainty Quantification for Segmentation with Multiple Annotations," **1** (2019).

[5] Pan, H., Feng, Y., Chen, Q., Meyer, C., and Feng, X., "Prostate segmentation from 3d mri using a two-stage model and variable-input based uncertainty measure," in [*Proceedings - International Symposium on Biomedical Imaging*], **2019-April**, 468–471 (3 2019).

[6] Ozdemir, O., Woodward, B., and Berlin, A. A., "Propagating Uncertainty in Multi-Stage Bayesian Convolutional Neural Networks with Application to Pulmonary Nodule Detection," (2017).

[7] Roy, A. G., Conjeti, S., Navab, N., and Wachinger, C., "Inherent brain segmentation quality control from fully convnet monte carlo sampling," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11070 LNCS**, 664–672 (2018).

[8] DeVries, T. and Taylor, G. W., "Leveraging Uncertainty Estimates for Predicting Segmentation Quality," (2018).

[9] Ronneberger, O., Fischer, P., and Thomas, B., "U-Net: Convolutional Networks for Biomedical Image Segmentation," in [*Medical Image Computing and Computer-Assisted Intervention (MICCAI)*], **9351**, 234–241 (2015).

[10] Lakshminarayanan, B., Pritzel, A., and Blundell, C., "Simple and scalable predictive uncertainty estimation using deep ensembles," in [*Advances in Neural Information Processing Systems*], **2017-Decem**, 6403–6414 (12 2017).

[11] Gal, Y. and Ghahramani, Z., "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," **48** (2015).

[12] Kendall, A., Badrinarayanan, V., and Cipolla, R., "Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding," (2015).

[13] Aberle, D. R., Adams, A. M., Berg, C. D., Black, W. C., Clapp, J. D., Fagerstrom, R. M., Gareen, I. F., Gatsonis, C., Marcus, P. M., and Sicks, J. R. D., "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England Journal of Medicine* **365**, 395–409 (8 2011).

[14] Martin, S. P., Hofmeister, J., Burgmeister, S., Orso, S., Mili, N., Guerrier, S., Victoria-Fesser, M. P., Soccal, P. M., Triponez, F., Karenovics, W., Mach, N., Depeursinge, A., Becker, C. D., Rampinelli, C., Summers, P., Müller, H., and Montet, X., "Identification of malignant lung nodules and reduction in false-positive findings by augmented intelligence: A radiomic study based on the NLST dataset," *Journal of Clinical Oncology* (submitted).

[15] Hendrycks, D. and Gimpel, K., "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," 1–12 (2016).

[16] Beers, A., Brown, J., Chang, K., Hoebel, K., Gerstner, E., Rosen, B., and Kalpathy-Cramer, J., "DeepNeuro: an open-source deep learning toolbox for neuroimaging," *arxiv* (8 2018).

[17] Kohl, S. A. A., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K. H., Eslami, S. M. A., Rezende, D. J., and Ronneberger, O., "A Probabilistic U-Net for Segmentation of Ambiguous Images," (6 2018).

[18] Wickstrøm, K., Kampffmeyer, M., and Jenssen, R., "Uncertainty and Interpretability in Convolutional Neural Networks for Semantic Segmentation of Colorectal Polyps," (2018).