Towards Integrating Epistemic Uncertainty Estimation into the Radiotherapy Workflow

Marvin Tom Teichmann[1(✉)], Manasi Datar[1], Lisa Kratzke[2], Fernando Vega[2], and Florin C. Ghesu[1]

marvin.teichmann@siemens-healthineers.com

[1] Digital Technology and Innovation, Siemens Healthineers, Erlangen, Germany
[2] Cancer Therapy Imaging, Varian, Siemens Healthineers, Forchheim, Germany

Abstract. The precision of contouring target structures and organs-at-risk (OAR) in radiotherapy planning is crucial for ensuring treatment efficacy and patient safety. Recent advancements in deep learning (DL) have significantly improved OAR contouring performance, yet the reliability of these models, especially in the presence of out-of-distribution (OOD) scenarios, remains a concern in clinical settings. This application study explores the integration of epistemic uncertainty estimation within the OAR contouring workflow to enable OOD detection in clinically relevant scenarios, using specifically compiled data. Furthermore, we introduce an advanced statistical method for OOD detection to enhance the methodological framework of uncertainty estimation. Our empirical evaluation demonstrates that epistemic uncertainty estimation is effective in identifying instances where model predictions are unreliable and may require an expert review. Notably, our approach achieves an AUC-ROC of 0.95 for OOD detection, with a specificity of 0.95 and a sensitivity of 0.92 for implant cases, underscoring its efficacy. This study addresses significant gaps in the current research landscape, such as the lack of ground truth for uncertainty estimation and limited empirical evaluations. This study addresses significant gaps in the current research landscape, such as the lack of ground truth for uncertainty estimation and limited empirical evaluations. Additionally, it provides a clinically relevant application of epistemic uncertainty estimation in an FDA-approved and widely used clinical solution for OAR segmentation from Varian, a Siemens Healthineers company, highlighting its practical benefits.

Keywords: Epistemic Uncertainty · Out-of-Distribution Detection · CT Segmentation · OAR contouring · Radiotherapy

The advent of precision medicine in radiotherapy has underscored the importance of accurate contouring of target structures and organs at risk (OAR) on computed tomography (CT) simulation scans. This critical step ensures that the dose objectives set by oncologists are met, directly influencing treatment outcomes. Recent advancements in DL have significantly enhanced the accuracy of contouring tasks, offering promising solutions to the challenges of manual delineation, which is time consuming, costly and suffers from inter-observer variability

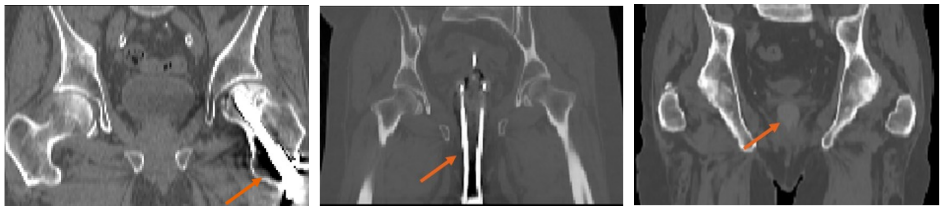
[5]. Despite these advancements, the reliability of DL models in clinical settings is often questioned, particularly when faced with OOD scenarios that were not represented in the training data. Such scenarios can lead to inaccuracies in contouring, posing significant risks to patient safety. In this application study we employ uncertainty estimation to address the challenges posed by OOD data. Specifically, we aim to evaluate the utility of uncertainty estimation for OOD detection, a critical aspect where traditional model confidences may fall short. Identification of OOD instances early in the radiotherapy workflow can help mitigate potential errors in the ensuing steps.

A recent review [18] examines the current research landscape in uncertainty estimation for medical imaging applications, highlighting its pivotal role in enhancing the reliability and trustworthiness of DL models in healthcare. Although it is an actively pursued area of research, the review identifies two major obstacles hindering progress and adaption: the lack of ground truth data and tasks for uncertainty estimation, and the limited empirical evaluations of uncertainty estimation methods, especially for real clinical applications. This study aims to address these aforementioned challenges within the realm of epistemic uncertainty estimation. To achieve this, we identify clinically relevant OOD scenarios for the OAR contouring task in radiotherapy. We then gather corresponding datasets and conduct empirical evaluations to assess the effectiveness of epistemic uncertainty estimation in those scenarios.

The contributions of our application study of uncertainty estimation for radiotherapy contouring are threefold: Firstly, we demonstrate the utility of uncertainty estimation for OOD detection by identifying clinically relevant use-cases and showing its effectiveness in mitigating associated risks. Secondly, we introduce a new OOD detection method using advanced statistical analysis for threshold estimation. Lastly, we aim to bridge the gap between research and clinical practice by showing how uncertainty estimation can provide quantifiable improvements for review of DL-contouring results in an FDA-approved and widely used pipeline for OAR segmentation by Varian, a Siemens Healthineers company. Our primary goal is to highlight the practical benefits of uncertainty estimation, and we aim to inspire further comparative evaluations on public datasets and advance the field of uncertainty quantification in medical imaging.

1 Our Datasets for Uncertainty Model Evaluation

In this application study, we evaluate the integration of uncertainty estimation and OOD detection within the framework of a clinical solution from Varian, a Siemens Healthineers company, designed for OAR contouring in radiotherapy. The study leverages a CT dataset comprising scans of the pelvic region, each resampled to an isometric resolution of 1 mm per voxel. Segmentation annotations for six organs – Bladder, Prostate, Rectum, Femoral Head Left, Femoral Head Right and Seminal Vesicles – were created in-house by a team consisting of radiologists, radiation oncologists and trained personal. The primary dataset



(a) (b) (c)

Fig. 1: Examples of OOD Scenarios in radiotherapy contouring: (a) Femur implants cause image artifacts. (b) Brachytherapy applicator devices distort the anatomy. (c) Hydrogel rectal spacer can significantly alter anatomical contours.

comprises 679 training and 20 *control* cases. In addition we collect three distinct OOD datasets:

1. Femur Implants Dataset: For this study, we created a dataset of 13 cases with femoral head implants, extracted from the original OAR contouring training set to simulate an engineering oversight scenario. As depicted in Figure 1(a), femoral head implants create artefacts that impact the anatomical appearance. This deliberate exclusion aims to evaluate the model's capability to identify OOD cases, where critical but potential rare data or patterns might not be part of the training set due to a selection bias.

2. Brachytherapy Dataset: This dataset encompasses 12 brachytherapy cases, where applicator devices induce anatomical distortions, especially in the rectum as illustrated in Figure 1(b). The clinical solution is currently not designed or certified for the use in Brachytherapy. The dataset represents a critical OOD scenario where a potential users are unaware of the OAR contouring limitations and unknowingly utilize it in an unsuitable context.

3. Hydrogel Rectal Spacer Dataset: We have collected a single case featuring a hydrogel rectal spacer, which significantly impacts anatomical contours, especially within the bladder and rectal regions, as illustrated in Figure 1(c). Similar to the brachytherapy cases, our clinical solution is not designed for scenarios involving hydrogel rectal spacers, marking it as another OOD case. There is a potential for practitioner oversight that could result in the model being applied in inappropriate contexts. Given its unique nature, this dataset is primarily utilized for qualitative evaluation.

These datasets, embodying various OOD scenarios, emphasize the critical role of uncertainty estimation in reinforcing system dependability and ensuring the safety of clinical practices.

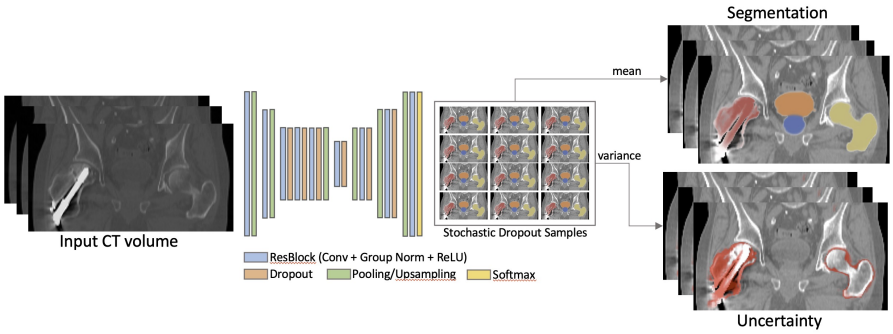


Fig. 2: Visualization of our uncertainty model for OAR contouring.

2 Our Method for Epistemic Uncertainty Estimation and OOD Detection

This section provides details of the architectural design, training protocol, and statistical methods for uncertainty quantification essential to integrate epistemic uncertainty estimation into the radiotherapy contouring workflow.

2.1 OAR Contouring: Architectural Design and Training Protocol

We approach OAR contouring as a 3D CT image segmentation task, processing 3D volumes ($H \times W \times T$) to produce segmentation maps with identical spatial dimensions. Our architecture, inspired by ConvNeXt [12] and illustrated in Figure 2, adopts a U-Net [14] style CNN with stages of ResBlocks – each comprising a 3D convolution, group normalization [17], and ReLU activation. Stages include downsampling and upsampling layers, with ResBlock counts of 1,1,3,1 and 1,1,1, respectively, and channel widths doubling after each downsampling and halving upon upsampling, starting from 16 channels. For training we minimize a softmax cross-entropy loss utilizing the AdamW optimizer [8,13], a batch size of 8, and a 1×10^{-3} initial learning rate, with a polynomial learning rate schedule [15], weight decay, and mixed-precision training. Augmentation techniques include random rotations ($\pm 5°$), rescaling ($\pm 20\,\%$) and gamma correction adjustments.

2.2 Methodology for Estimating Epistemic Uncertainty

We integrate a deep ensemble model [9] with Monte Carlo (MC) dropout [3], capitalizing on the strengths of both to augment uncertainty estimation within deep convolutional networks. This increases the sample size for uncertainty estimation, which is statistically advantageous offering a more robust measure. The selection of these methods is motivated by several key factors: 1) They specifically model epistemic uncertainty, aligning with our objective of performing OOD detection [7]; 2) Their theoretical foundations are solid and have been

well-received within the academic community, ensuring a rigorous approach to uncertainty estimation; 3) Their effectiveness in practical applications has been demonstrated repeatedly, validating their utility in real-world scenarios.

Deep Ensemble Configuration Our ensemble model consists of eight base learners, each trained on a subset of the training dataset with N samples to ensure diverse learning perspectives within the ensemble. To optimize the ensemble's coverage and diversity, the dataset is partitioned such that each base learner is exposed to at most $\frac{N}{2} + 1$ of cases, with every case being included in exactly four of the eight base models' training sets. We distribute the samples such that the overlap in training samples between any pair of base learners is limited to at most $\frac{N}{4} + 1$, thereby fostering diversity and maximizing the potential for disagreement among the learners within the specified constraints.

MC Dropout Implementation Our MC dropout implementation is inspired by the Bayesian SegNet framework [6]. Following the analysis of Kendall et al. [6], we only use dropout during the central stages of our architecture. However, our implementation diverges from Kendall's by inserting MC dropout layers after every ResBlock and not just at the end of each stage, thus more closely aligning with the foundational principles of MC dropout [9]. For all Dropouts we utilize 3D dropouts that remove an entire channel in the feature map, addressing the issue that adjacent voxels within feature maps are strongly correlated [16]. We determine the dropout rate using a hyperparameter search.

Inference and Uncertainty Quantification For inference, we generate 32 predictions per case by conducting four forward passes for each base learner with MC dropout enabled. An uncertainty heatmap is created by calculating the variance across these predictions for each voxel, yielding a segmentation map with 7 channels: one for each organ plus the background. Our post-processing utilizes morphological operations to filter out uncertainty responses at organ boundaries, attributed to annotation variance. An uncertainty score for each organ and case is derived by summing the remaining uncertainty responses within the respective organ channel of the uncertainty map. Uncertainty responses for the background are excluded from further analysis.

2.3 Statistical Method for OOD Detection

For statistical threshold estimation and OOD detection we utilize the Mahalanobis distance [11], traditionally applied to network feature distributions for OOD detection, which often necessitates network architecture modifications, such as flattening the final encoder layer [4] or averaging feature maps [2]. These adaptations can lead to potential feature collapse [10]. Contrary to these approaches, we directly apply the Mahalanobis distance to uncertainty scores derived from class-conditional distributions, offering a novel method that circumvents the need for architectural changes and avoids the risk of feature collapse.

As outlined in Section 2.2:*Deep Ensemble Configuration*, our training set is partitioned so each sample is used by only half of the base learners. This partitioning allows for the computation of conservative uncertainty scores u_i using the learners for which the sample was not in the training set. We approximate a class-conditional uncertainty distribution for the in-distribution (ID) training population by estimating M class-conditional Gaussian distributions $\mathcal{N}(\mu_m, \Sigma), m \in [1, M]$ over uncertainty scores u_i. Here, $\mu_m = \frac{1}{N}\sum_{i=m} u_i$ represents the class-wise means, and $\Sigma = \frac{1}{N}\sum_{m=1}^{M}\sum_{i=m}(u_i - \mu_m)(u_i - \mu_m)^T$ is the covariance matrix, capturing shared uncertainties across classes. For each test case with class-wise uncertainty scores z_i the Mahalanobis distance is now given as $\mathcal{DM}(z_i, \mathcal{N}(\mu m, \Sigma)) = \sqrt{\sum_{m=1}^{M}(z_i - \mu_m)^T \Sigma^{-1}(z_i - \mu_m)}$. An additional advantage is that the Mahalanobis distance follows a χ^2 distribution [1], with degrees of freedom equal to the number of features (foreground classes in this case), allowing for the computation of a critical value for OOD detection at a specific significance level without necessitating OOD or ID test samples.

3 Experimental Evaluation

In this section, we assess the effectiveness of our Bayesian epistemic uncertainty model in discriminating between ID and OOD cases in the OAR contouring workflow through qualitative analysis and quantitative assessment using the statistical methods discussed in Section 2.3.

3.1 Qualitative Analysis

Figure 3 showcases a representative case from each OOD dataset, illustrating confidence predictions, model uncertainty responses, and post-processed uncertainty for each scenario. Confidence levels are derived from the original deterministic contouring model, not the mean of the Bayesian model, and are visualized as soft-confidences using the alpha channel of the image overlay. A fully opaque pixel in the color of the corresponding class indicates a strong positive prediction by the deterministic model. Conversely, visibility of the background image signifies low confidence in any class, implying a high certainty of background rather than organ presence.

Femur Implants Figure 3(a) shows a case with an implant in the left femur head, where an analysis of confidence levels reveals high-confidence false positives and false negatives in the lower region of the left femur head. No low-confidence predictions are observed in this area, indicating that the model generates high-confidence predictions for this OOD case. Notably, our Bayesian model detects a significant uncertainty response in the lower part of the left femur head, accurately identifying the area where predictions are unreliable.

(a) Femur Implant: High uncertainty response at the femur implant.

(b) Brachytherapy: High uncertainty at rectum and prostate.

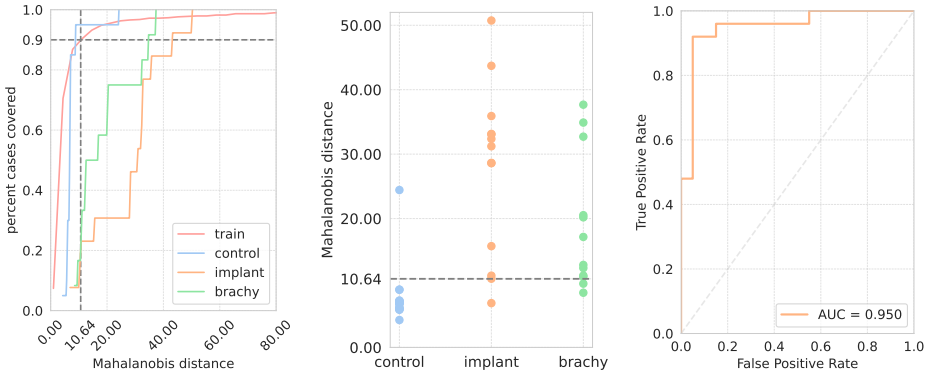
(c) Hydrogel Spacer: High uncertainty at the rectum.

Fig. 3: Visualization of results for OOD datasets with prediction (left), raw maximum uncertainties (center), processed maximum uncertainties (right). Maximum uncertainty is computed across all foreground classes (organs).

Brachytherapy Cases Figure 3(b) shows a case where a brachytherapy applicator deforms pelvic anatomy, leading to high-confidence false negative predictions for the bladder and prostate. This indicates that the model erroneously predicts with high confidence that these organs are absent. Crucially, our uncertainty model accurately flags areas where model predictions are unreliable due to anatomical distortion caused by the brachytherapy applicator.

Hydrogel Rectal Spacer Figure 3(c) illustrates a case with a hydrogel spacer inserted between the prostate and rectum, causing deformation of both organs. Despite these alterations, the model produces high-confidence predictions for the prostate, including high-confidence false positives in the spacer's vicinity, mistaking it for prostate tissue. Our approach effectively identifies regions where OAR predictions are affected by the anatomical changes due to the hydrogel spacer, evidencing the model's ability to detect uncertainty in scenarios where organ contours are modified by medical interventions.

In summary, our qualitative evaluation shows that model confidences alone do not paint the full picture, especially in OOD scenarios where confidence lev-



(a) (b) (c)

Fig. 4: Statistical analysis: (a) Cumulative class-conditional Mahalanobis distance curves, x-axis is truncated. (b) Scatterplot of Mahalanobis distance scores for each dataset including OOD threshold. (c) ROC curve for OOD detection.

els often misrepresent model certainty and fail to provide a reliable indicator of model performance. Conversely, our Bayesian uncertainty approach offers a more accurate depiction of model reliability, effectively identifying areas where predictions are uncertain.

3.2 Quantitative Analysis

We use the class-conditional uncertainty distribution based in the training data as discussed in Section 2.3 and estimate the Mahalanobis distance from this distribution for each sample in the ID (training, control) and OOD (implant, brachy) datasets. Using the train distribution we estimate an OOD threshold of 10.64 utilizing the χ^2 table [1] for an application-specific critical value of 0.9 and degrees of freedom equal to 6 (number of foreground classes). Figure 4(a) illustrates the threshold selection. Note that only the training set is required for the threshold estimation, no OOD or control samples are utilized.

Figure 4(b) illustrates the OOD-detection efficacy of our threshold on the ID and OOD test sets. The threshold accurately classifies 95.0% of control cases as ID (Specificity), and identifies 92.3% of Implant cases and 83.3% of Brachy cases as OOD (Sensitivity). The choice of a critical value, and thus the Sensitivity-Specificity trade-off, may vary based on specific application needs. For a comprehensive evaluation, Figure 4(c) showcases our method's threshold-independent performance through an ROC curve, derived from class-conditional Mahalanobis distances across the test dataset. With ID (control) class = 0 and OOD (brachy, implant) class = 1, we achieve an AUC-ROC of 0.95. This demonstrates that our distance-based metric, derived from the underlying multivariate uncertainty distribution, effectively discriminates OOD samples with statistical significance.

4 Conclusion & Future Work

In this study, we have explored the integration of epistemic uncertainty estimation within the radiotherapy contouring workflow, particularly focusing on enhancing the detection of OOD scenarios. Our investigation was motivated by the critical need to improve the safety and reliability of OAR models in radiotherapy, where the contouring precision for target structures and OAR is paramount for effective treatment planning and the safety of the patient. We were able to demonstrate the effectiveness and utility of epistemic uncertainty estimation in identifying instances where model predictions are unreliable. Our approach achieved an AUC-ROC of 0.95 for OOD detection, with a specificity of 0.95 and a sensitivity of 0.92 for implant cases. These findings highlight the potential of uncertainty estimation as an early warning system for cases requiring additional attention during requiring expert review. By doing so we also address the previously identified gaps in the current research landscape [18], namely the lack of ground truth data and tasks as well as the need for empirical evaluations.

Looking forward, our study lays the groundwork for further research into the integration of uncertainty analysis into the radiotherapy workflow. The promising results from our initial feasibility study suggest that epistemic uncertainty estimation can significantly enhance the accuracy, safety, and clinical applicability of deep learning models in radiotherapy planning. Future work will focus on expanding the range of OOD scenarios explored, refining our methodology, and conducting empirical comparisons. Additionally, we encourage the academic community to develop benchmarks and similar tasks based on public data, which can be used for quantitative comparisons and empirical evaluations of various methods. There remains a significant gap in the comparative analysis of uncertainty estimation methods, as highlighted in [18]. Our work provides meaningful, clinically relevant tasks that the community can use to quantitatively evaluate methods. We hope that our work and results highlight the importance of further research in the space of uncertainty estimation for medical imaging, driving advancements in this critical area.

Disclaimer The information in this paper is based on research results that are not commercially available. Future commercial availability cannot be guaranteed.

Disclosure of Interests. The authors are employed by Siemens Healthineers, and the research was fully funded by Siemens Healthineers. The authors have no further competing interests to declare that are relevant to the content of this article.

References

1. Brereton, R.G.: The chi squared and multinormal distributions. Journal of Chemometrics **29**(1), 9–12 (2015). https://doi.org/https://doi.org/10.1002/cem.2680, https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.2680

2. Calli, E., Van Ginneken, B., Sogancioglu, E., Murphy, K.: Frodo: An in-depth analysis of a system to reject outlier samples from a trained neural network. IEEE Transactions on Medical Imaging **42**(4), 971–981 (2023). https://doi.org/10.1109/TMI.2022.3221898
3. Gal, Y., Ghahramani, Z.: Bayesian convolutional neural networks with bernoulli approximate variational inference. arXiv preprint arXiv:1506.02158 (2015)
4. González, C., Gotkowski, K., Fuchs, M., Bucher, A., Dadras, A., Fischbach, R., Kaltenborn, I.J., Mukhopadhyay, A.: Distance-based detection of out-of-distribution silent failures for covid-19 lung lesion segmentation. Medical Image Analysis **82**, 102596 (2022). https://doi.org/https://doi.org/10.1016/j.media.2022.102596, https://www.sciencedirect.com/science/article/pii/S1361841522002298
5. Joskowicz, L., Cohen, D., Caplan, N., Sosna, J.: Inter-observer variability of manual contour delineation of structures in ct. European radiology **29**, 1391–1399 (2019)
6. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680 (2015)
7. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
10. Lambert, B., Forbes, F., Doyle, S., Dojat, M.: Multi-layer aggregation as a key to feature-based ood detection. In: Sudre, C.H., Baumgartner, C.F., Dalca, A., Mehta, R., Qin, C., Wells, W.M. (eds.) Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 104–114. Springer Nature Switzerland, Cham (2023)
11. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems **31** (2018)
12. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
14. 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)
15. Teichmann, M.T., Cipolla, R.: Convolutional crfs for semantic segmentation. British Machine Vision Conference (BMVC) (2018)
16. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 648–656 (2015)
17. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
18. Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., Fu, H.: A review of uncertainty estimation and its application in medical imaging. arXiv preprint arXiv:2302.08119 (2023)