On the relationship between calibrated predictors and unbiased volume estimation

Teodora Popordanoska, Jeroen Bertels, Dirk Vandermeulen, Frederik Maes, and Matthew B. Blaschko

Center for Processing Speech and Images, Dept. ESAT, KU Leuven, Belgium
teodora.popordanoska@kuleuven.be

Abstract. Machine learning driven medical image segmentation has become standard in medical image analysis. However, deep learning models are prone to overconfident predictions. This has lead to a renewed focus on *calibrated predictions* in the medical imaging and broader machine learning communities. Calibrated predictions are estimates of the probability of a label that correspond to the true expected value of the label conditioned on the confidence. Such calibrated predictions have utility in a range of medical imaging applications, including surgical planning under uncertainty and active learning systems. At the same time it is often an accurate *volume measurement* that is of real importance for many medical applications. This work investigates the relationship between model calibration and volume estimation. We demonstrate both mathematically and empirically that if the predictor is calibrated *per image*, we can obtain the correct volume by taking an expectation of the probability scores per pixel/voxel of the image. Furthermore, we show that linear combinations of calibrated classifiers preserve volume estimation, but do not preserve calibration. Therefore, we conclude that having a calibrated predictor is a sufficient, but not necessary condition for obtaining an unbiased estimate of the volume. We validate our theoretical findings empirically on a collection of 18 different (calibrated) training strategies on the tasks of glioma volume estimation on BraTS 2018, and ischemic stroke lesion volume estimation on ISLES 2018 datasets.

Keywords: Calibration · Uncertainty · Volume · Segmentation

1 Introduction

In recent years the segmentation performance of CNNs improved dramatically. Despite these improvements, the adoption of automated segmentation systems into clinical routine is rather slow. Having calibrated model predictions would foster this relationship by providing the clinician with valuable information on failure detection or when manual intervention is needed [15]. Reasoning under uncertainty is a central part of surgical planning [11], and uncertainty estimates are central to many active learning frameworks [28]. With this in mind, current research orients towards the design of automated segmentation systems that provide a measure of confidence alongside its predictions.

The empirical findings that modern CNNs are poorly calibrated [13] stimulated a variety of research to improve model calibration, including deep ensembles [18], Bayesian NNs [22] and MC dropout [10]. The ability to deliver realistic segmentations would take calibration even further [16, 3]. Remarkably, a large part of ongoing research stands orthogonal by design. Choosing a loss function that is consistent with a certain target metric [9] often means the outputs cannot be interpreted as voxel-wise probabilities [4], let alone that the predictions would be calibrated. Nonetheless, post-hoc calibration might come to the rescue and has proven to be competitive to training-time calibration [24].

While the automated segmentation is an important task in its own right, it is mostly used as an intermediary for calculating certain biomarkers. In that respect, volume is by far the most important biomarker in medical imaging. For example, tumor volume is a basic and specific response predictor in radiotherapy [7] and the volume of an acute stroke lesion can be used to decide on the type of endovascular treatment [12]. When the CNN outputs voxel-wise probabilities, they can be propagated to produce volumetric uncertainty [8]. However, similar to its effect on calibration, the loss function determines how to calculate volume, and thus when the former is applicable [4].

This work bridges the gap between model calibration and volume estimation. There will be theoretical grounds that calibration error bounds volume error. This relationship is confirmed in an empirical validation on BraTS 2018 [20, 1, 2] and ISLES 2018 [27]. Furthermore, there is a clear empirical correlation between calibration error and volume error, and between object size and volume error. As a result, this work acknowledges and encourages research towards calibrated systems. Such systems not only provide additional robustness; they will also produce correct volume estimates.

2 The relationship between calibration and volume bias

In this section, we demonstrate that calibrated uncertainty estimates are intimately related to unbiased volume estimates. We develop novel mathematical results showing that calibration error upper bounds the absolute value of the volume bias (Proposition 1), which implies that as the calibration error goes to zero, the resulting function has unbiased volume estimates. We further show that unbiased volume estimates do not imply that the classifier is calibrated, and that solely enforcing an unbiased classifier does not result in a calibrated classifier (Proposition 2 and Corollary 2). This motivates our subsequent experimental study where we measure the empirical relationship between calibration error and volume bias in Sections 3.

Definition 1 (Volume bias [4]). *Let f be a function that predicts from an image/tomography x the probability of each pixel/voxel $\{y_i\}_{i=1}^{p}$ belonging to a given class. The* volume bias *of f is:*

$$\text{Bias}(f) := \mathbb{E}_{(x,y)\sim P}\left[f(x) - y\right]. \tag{1}$$

Definition 2 (Calibration error [21, 17, 26]). *The calibration error of* $f : \mathcal{X} \to [0, 1]$ *is:*

$$\mathrm{CE}(f) = \mathbb{E}_{(x,y)\sim P}\left[\left|\mathbb{E}_{(x,y)\sim P}\left[[y = 1] \mid f(x)\right] - f(x)\right|\right] \tag{2}$$

In plain English: A classifier is calibrated if its confidence score is equal to the probability of the prediction being correct. Note that this definition is specific to a binary classification setting. The extension of binary calibration methods to multiple classes is usually done by reducing the problem of multiclass classification to K one-vs.-all binary problems [13]. The so-called *marginal CE* [17, Definition 2.4]) is then measured as an average of the per-class CEs.

Proposition 1. *The absolute value of dataset (respectively volume) bias is upper bounded by dataset (respectively volume) calibration error:*

$$\mathrm{CE}(f) \geq |\,\mathrm{Bias}(f)|. \tag{3}$$

Proof.

$$|\,\mathrm{Bias}(f)| = \left|\mathbb{E}_{(x,y)\sim P}\left[y - f(x)\right]\right| \tag{4}$$

$$= \left|\underbrace{\mathbb{E}\left[y - \mathbb{E}\left[[y = 1] \mid f(x)\right]\right]}_{=0} + \mathbb{E}_{(x,y)\sim P}\left[\mathbb{E}\left[[y = 1] \mid f(x)\right] - f(x)\right]\right| \tag{5}$$

$$\leq \underbrace{\mathbb{E}_{(x,y)\sim P}\left[\left|\mathbb{E}\left[[y = 1] \mid f(x)\right] - f(x)\right|\right]}_{=\mathrm{CE}(f)}. \tag{6}$$

Focusing on the first term of the right hand side of (5),

$$\mathbb{E}\left[y - \mathbb{E}[[y = 1] \mid f(x)]\right] = \mathbb{E}[y] - \underbrace{\mathbb{E}[\mathbb{E}\left[[y = 1] \mid f(x)\right]]}_{=\mathbb{E}[y]}. \tag{7}$$

In the second term of the r.h.s., we may take the expectation with respect to $f(x)$ in place of x as f is a deterministic function. This term is therefore also equal to $\mathbb{E}[y]$ by the law of total expectation. Finally, the inequality in (6) is obtained due to the convexity of the absolute value and by application of Jensen's inequality. □

We note that Proposition 1 holds for all problem settings and class distributions. In multiclass settings, for each individual class we can obtain a bound by measuring the bias and the CE per class.

Corollary 1. $\mathrm{CE}(f) = 0$ *implies that* f *yields unbiased volume estimates.*

Proposition 2. $\mathrm{Bias}(f) = 0$ *does not imply that* $\mathrm{CE}(f) = 0$.

Proof. Consider the following example. Let the dataset consist of 100 positive and 200 negative points. Let f_1 and f_2 be binary classifiers that rank 1/4 of the

negative points with a score of zero, and the remaining negative points with a score of 0.25. Let f_1 rank the first half of the positive points with a score of one and the second half with a score of 0.25, and f_2 vice-versa. In this case, $\text{CE}(f_1) = 0$ and $\text{CE}(f_2) = 0$, and therefore by Corollary 1, f_1 and f_2 are both unbiased estimates of the volume. Let f_3 be a classifier that performs a linear combination (e.g. an average) of the scores of f_1 and f_2. We note that a convex combination of unbiased estimators is unbiased [19] and therefore f_3 is unbiased. Even though f_3 has a perfect accuracy and $\text{Bias}(f_3) = 0$, the scores are no longer calibrated, i.e., $\text{CE}(f_3) \neq 0$. □

Corollary 2. *There exists no multiplicative bound of the form* $\text{CE}(f) \leq \gamma |\,\text{Bias}(f)|$ *for some finite* $\gamma > 0$ *(cf. [9, Definition 2.2]).*

Thus we see that control over $\text{CE}(f)$ minimizes an upper bound on $|\,\text{Bias}(f)|$, but the converse is not true. There are several implications of these theoretical results for the design of medical image analysis systems: (i) Optimizing calibration error per-subject is an attractive method to simultaneously control the bias of the volume estimate, but we need to empirically validate if bias and CE are correlated in practice as we only know *a priori* that one bounds the other; and (ii) optimizing volume bias alone (e.g. by empirical risk minimization) does not automatically give us the additional benefits of calibrated uncertainty estimates, and does not even provide a multiplicative bound on how poor the calibration error could be. We consequently empirically evaluate the relationship between bias and calibration error in the remainder of this work.

3 Empirical setup

The empirical validation of our theoretical results will be performed by analyzing two segmentation tasks, each requiring a different distribution in the predicted confidences, with multiple different models, each trained with respect to a different loss function and subject to different post-hoc calibration strategies.

Tasks The data from two publicly available medical datasets is used and two segmentation tasks are formulated as follows: (i) Whole tumor segmentation using the BraTS 2018 [20, 1, 2] (BR18) dataset. BR18 contains 285 multi-modal MR volumes with accompanying manual tumor delineations. Due to a rather low inter/intra-rater variability [20] the voxel-wise confidences will be distributed towards the high-confidence ranges; (ii) Ischemic core segmentation using the ISLES 2018 [27] (IS18) dataset. IS18 contains data from 94 CT perfusion scans with manual delineations of the ischemic infarctions on co-registered DWI MR imaging. The identification of the ischemic infarction on CT perfusion data is generally considered non-trivial [6]. This means that a rather high intra/inter-rater variability is to be expected, which in turn will distribute the voxel-wise confidences across the entire range. For both datasets there was a five-fold split of the data identical to [4, 24].

Models For the two former tasks the pre-trained and publicly available models from [24] are used. They investigated the effects of the loss function in combination with a multitude of different post-hoc calibration strategies on the Dice

score and model calibration, but without any consideration to volume estimates or volume bias. Their base model shares a U-Net [23] CNN architecture similar to [14]. The three loss functions for the initial training were: (i) cross-entropy (CrE); (ii) soft-Dice (SD); and (iii) a combination of CrE pre-training with SD fine-tuning (CrE-SD). In addition, these base models were calibrated using different post-hoc calibration strategies: (i) Platt scaling and its variants (auxiliary network and fine-tuning); and (ii) two Monte Carlo (MC) dropout methods (MC-Dropout and MC-Center) with different positioning of the dropout layers. For further details on the exact training procedures the reader is referred to [24]. Nevertheless, it is important to note that the initial training and the post-hoc calibration was done on the training sets, and thus the predictions on the validation sets may be aggregated for further testing.

Bias and ECE The Bias is calculated by direct implementation of Definition 1, i.e. the expectation of the probability scores per voxel. The CE from Definition 2 is a theoretical quantity that in practice is approximated by a binned estimator of the expected calibration error (ECE) (Equation (3) from [24]). The ECE ranges from 0 to 1, with lower values representing better calibration. We use 20 bins for binning the CNN outputs. Following the example of [24, 15], we only consider voxels within the skull-stripped brain/lesion and report the mean per-volume Bias and ECE, as being more clinically relevant versions opposed to their dataset-level variants.

Code The source code is available at `https://github.com/tpopordanoska/calibration_and_bias`.

4 Results and Discussion

Fig. 1 shows scatter plots of ECE and Bias for the base model and a calibrated model with the fine-tuning strategy for BR18 and IS18. We can confirm that, as predicted by Proposition 1, all the points (volumes measured in ml) lie in-between the lines ECE $= \pm$ Bias. We note further that the correlation between ECE and Bias is strictly higher for a calibrated model compared to the base model. There are many volumes for which ECE $=|$ Bias $|$ holds, which supports the findings in [15] where per-volume calibration tends to be off, either resulting in a complete under- or over-estimation. This is also visible in Fig. 3, when the calibration curve lies below or above the unity line ECE $=|$ Bias $|$.

The ECE and Bias for all 18 models are visually presented in a scatter plot in Fig. 2. Analogously to [24], we find that for the BR18 data only the models trained with SD are on the Pareto front. However, contrary to their result that MC methods are Pareto dominated if optimizing the Dice score is of interest, we observe that the MC-Decoder calibration strategy is Pareto-efficient for the clinical application of measuring volume. For IS18, the model trained with CrE and calibrated with the MC strategy Pareto dominates all the rest. Therefore, we conclude that when selecting a model for volume estimation, prioritization of lowering the ECE over choosing the right training loss is advised. This somewhat refines the findings in [4] where CrE outperformed SD variants in terms of volume

bias on a dataset level. This is further exemplified in Fig. 3 which shows visually on selected slices of different-sized volumes that the ECE is a more reliable predictor of the Bias than the loss function used during training.

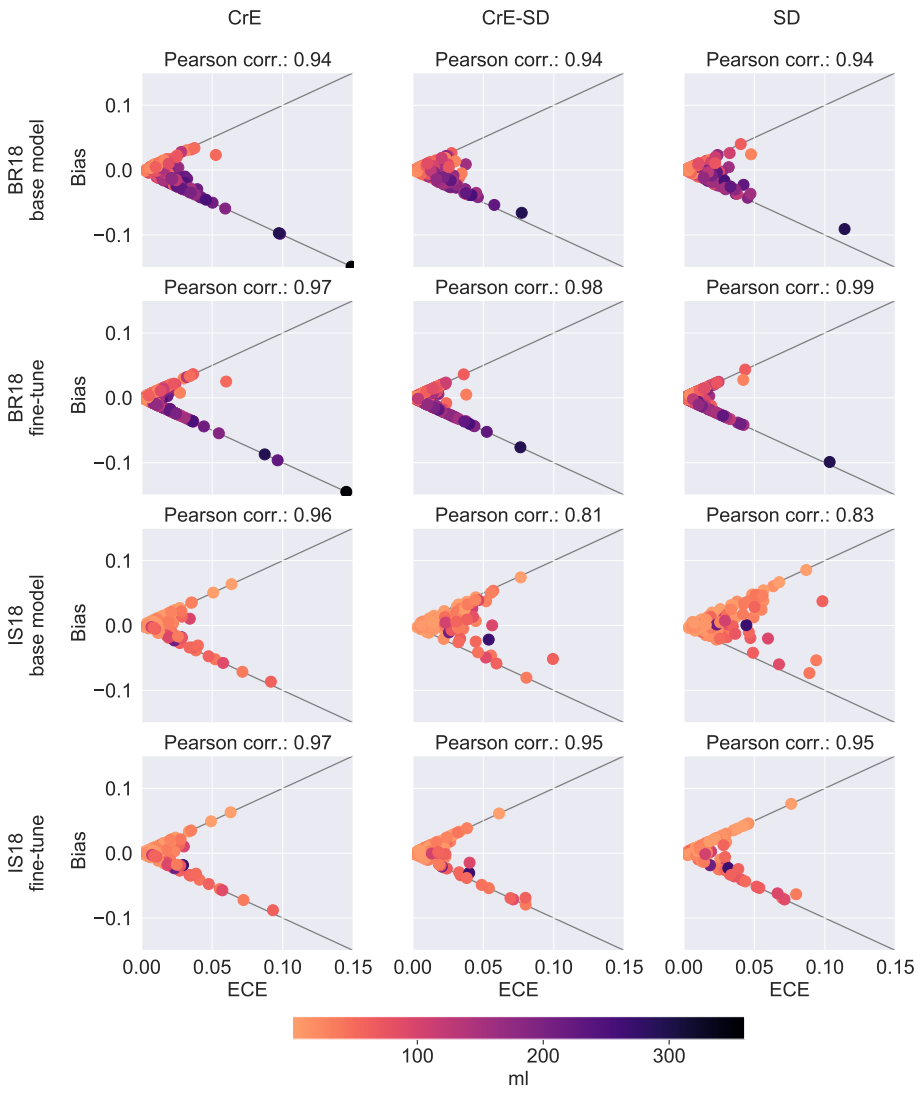

Fig. 1: Scatter plots on BR18 (top two rows) and IS18 (bottom two rows), color-coded by tumor/lesion size (in ml). Every point in the plot represents an image. The Pearson correlation between per-volume ECE and absolute per-volume bias is shown above the plots. Note that the Bias is calculated in voxels, while the color-coding has the units converted to ml.

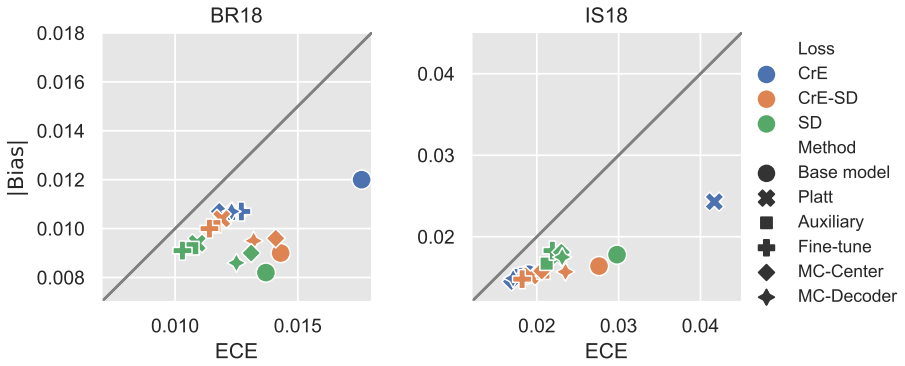

Fig. 2: Scatter plots of ECE versus |Bias| for all combinations of loss functions and calibration methods. The Pearson correlation between |Bias| (computed as mean of absolute per-volume biases) and ECE (mean per-volume ECE) is 0.30 $\pm$ 0.21 for BR18 and 0.91 $\pm$ 0.04 for IS18.

As an additional experiment on the difference between datasets with different levels of inter-rater variability, we calculated correlations between the mean absolute per-volume Bias and mean per-volume ECE for the 18 settings on the BR18 dataset separated into high grade (HGG) and low grade glioma (LGG). The Pearson correlation is 0.15 $\pm$ 0.23 for HGG and 0.39 $\pm$ 0.20 for LGG. Compared to the analogous result in the caption of Figure 2, we observe again that the correlation is higher for the data with higher uncertainty (LGG and IS18).

Table 1 shows the Pearson correlation coefficients with error bars [5] between the per-volume bias and the tumor/lesion size for both datasets. Negative correlations are also visible in the color-coded volumes in Fig. 1. All models have a tendency to underestimate large volumes (negative bias) and overestimate small volumes (positive bias). This trend was also observed in [4, 25] where simple post-hoc training-set regression was used for recalibration. We further observe that the correlation is stronger for the volumes in the BR18 than IS18 data. However, there is no consistent finding that holds for both datasets (e.g. the effect on the correlation is not influenced by the loss or post-calibration method). Additionally, we calculated the Spearman ρ and Kendall τ coefficients for all settings. In all cases there are significant non-zero correlations and the median values for BR18 are -0.43 (ρ) and -0.29 (τ), and for IS18 -0.43 (ρ) and -0.30 (τ).

Limitations We showed theoretically and empirically that the CE is an upper bound for the Bias, and that the Bias does not provide a bound on CE. Nonetheless, we did not properly characterize when CE may be strictly larger than this quantity and whether there is a meaningful interpretation of CE $-$ |Bias|. Furthermore, we proved that a convex combination of calibrated classifiers does not preserve calibration (although it preserves unbiasedness), but we did not explore ways to combine calibrated models to obtain a calibrated predictor.

Table 1: Pearson correlation coefficients between per-volume bias and volume size for BR18 and IS18.

loss→	CrE	CrE-SD	SD	CrE	CrE-SD	SD
method ↓	BR18			IS18		
base model	-0.73 ± 0.03	-0.53 ± 0.04	-0.48 ± 0.05	-0.39 ± 0.09	-0.23 ± 0.10	-0.20 ± 0.10
Platt	-0.58 ± 0.04	-0.67 ± 0.03	-0.62 ± 0.04	-0.50 ± 0.08	-0.39 ± 0.09	-0.48 ± 0.08
auxiliary	-0.57 ± 0.04	-0.65 ± 0.03	-0.61 ± 0.04	-0.32 ± 0.09	-0.31 ± 0.09	-0.43 ± 0.08
fine-tune	-0.59 ± 0.04	-0.62 ± 0.04	-0.55 ± 0.04	-0.37 ± 0.09	-0.36 ± 0.09	-0.49 ± 0.08
MC-Decoder	-0.55 ± 0.04	-0.47 ± 0.05	-0.42 ± 0.05	-0.23 ± 0.10	-0.27 ± 0.10	-0.36 ± 0.09
MC-Center	-0.53 ± 0.04	-0.49 ± 0.04	-0.49 ± 0.04	-0.27 ± 0.10	-0.24 ± 0.10	-0.39 ± 0.09

Designing a calibrated estimator is a non-trivial task and the field of confidence calibration is an active area of research. However, investing resources to find out which calibration strategy works best for the problem at hand is of high importance, especially in medical applications.

Finally, we wish to emphasize the distinction between dataset and per-volume ECE. Often in clinical settings, the information about per-subject calibration is of higher importance. A zero dataset level ECE implies zero dataset bias, however, the subject-level biases may very well be non-zero. This subject is treated less frequently in the calibration literature.

5 Conclusions

In this work, it was shown that the importance of confidence calibration goes beyond using the predicted voxel-wise confidences for clinical guidance and robustness. More specifically, there is theoretical and empirical evidence that calibration error bounds the error of volume estimation, which still is one of the most relevant biomarkers calculated further downstream. Since the converse relationship does not hold, the direct optimization of calibration error is to be preferred over the optimization of volume bias alone.

Acknowledgments

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. J.B. is part of NEXIS (`www.nexis-project.eu`), a project that has received funding from the European Union's Horizon 2020 Research and Innovations Programme (Grant Agreement #780026).

References

1. Bakas, S., Akbari, H., Sotiras, A., et al.: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**(March), 1–13 (2017)

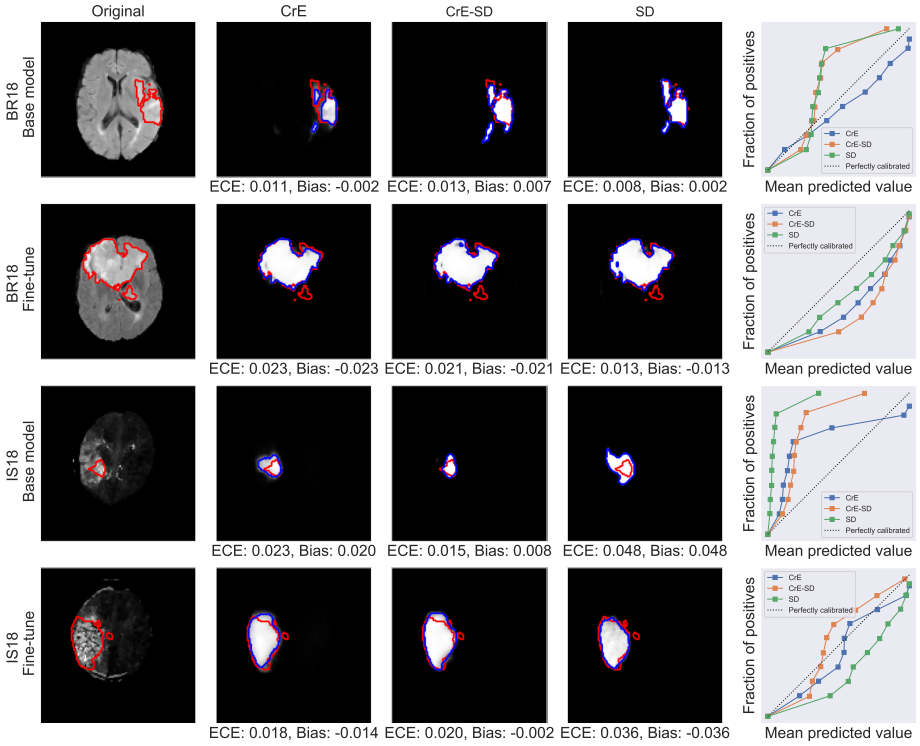

Fig. 3: Qualitative examples of the predictions from a base model and a calibrated model with fine-tuning on BR18 (top two rows) and IS18 (bottom two rows). The red line represents the delineation of the ground truth. The predicted delineations after thresholding at 0.5 are overlayed in blue. The ECE and the bias (shown below the plot) are calculated for the selected slice. The last column shows the calibration curves for the volume. The images are chosen to be representative of different sizes of tumors/lesions and the middle slice of the volume is shown.

2. Bakas, S., Reyes, M., Jakab, A., et al.: Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge (2018)
3. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., et al.: Phiseg: Capturing uncertainty in medical image segmentation. In: MICCAI. pp. 119–127 (2019)
4. Bertels, J., Robben, D., Vandermeulen, D., Suetens, P.: Theoretical analysis and experimental validation of volume bias of soft dice optimized segmentation maps in the context of inherent uncertainty. Medical Image Analysis **67**, 101833 (2021)
5. Bowley, A.L.: The standard deviation of the correlation coefficient. Journal of the American Statistical Association **23**(161), 31–34 (1928)
6. Demeestere, J., Garcia-Esperon, C., Garcia-Bermejo, P., et al.: Evaluation of hyperacute infarct volume using ASPECTS and brain CT perfusion core volume. Neurology **88**(24), 2248–2253 (2017)

7. Dubben, H.H., Thames, H.D., Beck-Bornholdt, H.P.: Tumor volume: a basic and specific response predictor in radiotherapy. Radiotherapy and Oncology **47**(2), 167–174 (1998)
8. Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., Cardoso, M.J.: Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI. pp. 691–699 (2018)
9. Eelbode, T., Bertels, J., Berman, M., et al.: Optimization for medical image segmentation: Theory and practice when evaluating with Dice score or Jaccard index. IEEE Transactions on Medical Imaging **39**(11), 3679–3690 (2020)
10. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. pp. 1050–1059 (2016)
11. Gillmann, C., Maack, R.G., Post, T., Wischgoll, T., Hagen, H.: An uncertainty-aware workflow for keyhole surgery planning using hierarchical image semantics. Visual Informatics **2**(1), 26–36 (2018)
12. Goyal, M., Menon, B.K., Zwam, W.H.V., et al.: Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. The Lancet **387**(10029), 1723–1731 (2016)
13. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. vol. 70, pp. 1321–1330 (2017)
14. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: No new-net. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 234–244. Springer (2019)
15. Jungo, A., Balsiger, F., Reyes, M.: Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Frontiers in Neuroscience **14**, 282 (2020)
16. Kohl, S.A., Romera-Paredes, B., Meyer, C., et al.: A probabilistic U-net for segmentation of ambiguous images. In: Advances in Neural Information Processing Systems, pp. 6965–6975 (2018)
17. Kumar, A., Liang, P.S., Ma, T.: Verified uncertainty calibration. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 3792–3803 (2019)
18. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, pp. 6403–6414 (2017)
19. Lee, A.J.: U-Statistics: Theory and Practice. Taylor & Francis (1990)
20. Menze, B.H., Jakab, A., Bauer, S., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging (2015)
21. Naeini, M.P., Cooper, G.F., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. pp. 2901–2907 (2015)
22. Neal, R.M.: Bayesian learning for neural networks. Springer (2012)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
24. Rousseau, A.J., Becker, T., Bertels, J., Blaschko, M., Valkenborg, D.: Post training uncertainty calibration of deep networks for medical image segmentation. In: ISBI (2021)

25. Tilborghs, S., Maes, F.: Left ventricular parameter regression from deep feature maps of a jointly trained segmentation cnn. In: Pop, M., Sermesant, M., Camara, O., Zhuang, X., Li, S., Young, A., Mansi, T., Suinesiaputra, A. (eds.) Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges. pp. 395–404. Springer (2020)
26. Wenger, J., Kjellström, H., Triebel, R.: Non-parametric calibration for classification. In: International Conference on Artificial Intelligence and Statistics. pp. 178–190 (2020)
27. Winzeck, S., Hakim, A., McKinley, R., et al.: ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI. Frontiers in Neurology **9**, 679 (2018)
28. Wu, J., Ruan, S., Lian, C., et al.: Active learning with noise modeling for medical image annotation. In: ISBI. pp. 298–301 (2018)

On the relationship between calibrated predictors and unbiased volume estimation: Supplementary material

Teodora Popordanoska, Jeroen Bertels, Dirk Vandermeulen, Frederik Maes, and Matthew B. Blaschko

Center for Processing Speech and Images, Dept. ESAT, KU Leuven, Belgium
teodora.popordanoska@kuleuven.be

Table 1: Summary of mean per-volume ECE and mean absolute per-volume Bias for all combinations of loss functions and methods on BR18.

loss→	CrE	CrE-SD	SD	CrE	CrE-SD	SD
method ↓	| Bias |			ECE		
base model	0.0120 ± .0009	0.0090 ± .0008	0.0082 ± .0008	0.0176 ± .0008	0.0143 ± .0008	0.0137 ± .0008
Platt	0.0107 ± .0008	0.0104 ± .0008	0.0094 ± .0008	0.0125 ± .0008	0.0119 ± .0008	0.0109 ± .0008
auxiliary	0.0106 ± .0008	0.0101 ± .0008	0.0092 ± .0008	0.0122 ± .0008	0.0116 ± .0008	0.0107 ± .0008
fine-tune	0.0107 ± .0008	0.0100 ± .0008	0.0091 ± .0008	0.0127 ± .0008	0.0114 ± .0008	0.0103 ± .0008
MC-Decoder	0.0107 ± .0007	0.0095 ± .0008	0.0086 ± .0008	0.0123 ± .0007	0.0132 ± .0009	0.0125 ± .0008
MC-Center	0.0107 ± .0008	0.0096 ± .0009	0.0090 ± .0008	0.0118 ± .0008	0.0141 ± .0009	0.0131 ± .0009

Table 2: Summary of mean per-volume ECE and mean absolute per-volume Bias for all combinations of loss functions and methods on IS18.

loss→	CrE	CrE-SD	SD	CrE	CrE-SD	SD
method ↓	| Bias |			ECE		
base model	0.0154 ± .0017	0.0164 ± .0018	0.0178 ± .0019	0.0190 ± .0016	0.0276 ± .0019	0.0298 ± .0022
Platt	0.0243 ± .0022	0.0152 ± .0018	0.0175 ± .0019	0.0417 ± .0031	0.0192 ± .0017	0.0225 ± .0018
auxiliary	0.0153 ± .0017	0.0156 ± .0018	0.0167 ± .0018	0.0180 ± .0017	0.0208 ± .0017	0.0212 ± .0018
fine-tune	0.0151 ± .0017	0.0148 ± .0018	0.0183 ± .0018	0.0181 ± .0017	0.0182 ± .0018	0.0219 ± .0018
MC-Decoder	0.0149 ± .0018	0.0157 ± .0018	0.0175 ± .0017	0.0171 ± .0018	0.0235 ± .0018	0.0231 ± .0020
MC-Center	0.0145 ± .0016	0.0158 ± .0017	0.0181 ± .0018	0.0169 ± .0016	0.0206 ± .0016	0.0230 ± .0019

Table 3: Tabulated list of parameters used in various calibration strategies.

parameter→	batch size	initial learning rate	epochs
base model	2 for BR18; 4 for IS18	10^{-3}	until convergence
Platt	64 z-slices	$5 \cdot 10^{-3}$	max 50
auxiliary	64 z-slices	$5 \cdot 10^{-3}$	max 50
fine-tune	2 3D volumes	10^{-3} for SD; 10^{-4} for CrE	max 50
MC methods	2 3D volumes	best of $\{10^{-3}, 10^{-4}, 10^{-5}\}$	max 50

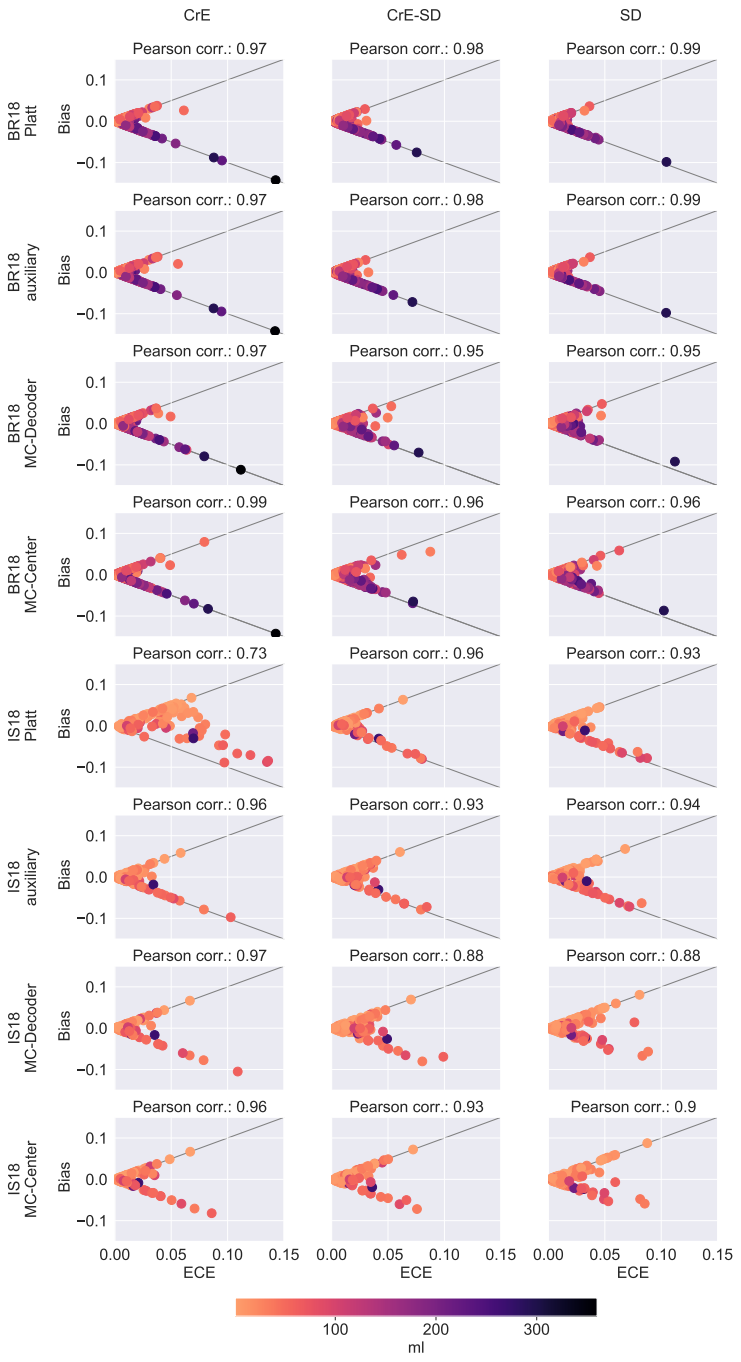

Fig. 1: Scatter plots on BR18 (top four rows) and IS18 (bottom four rows), color-coded by tumor/lesion size (in ml) for the remaining four calibration strategies. Every point in the plot represents an image. The Pearson correlation between per-volume ECE and absolute per-volume bias is shown above the plots.