Recalibration of Aleatoric and Epistemic Regression Uncertainty in Medical Imaging

Max-Heinrich Laves   max-heinrich.laves@tuhh.de
Institute of Medical Technology and Intelligent Systems, Hamburg University of Technology, Hamburg, Germany
Institute of Mechatronic Systems, Leibniz Universität Hannover, Hannover, Germany

Sontje Ihler   ihler@imes.uni-hannover.de
Institute of Mechatronic Systems, Leibniz Universität Hannover, Hannover, Germany

Jacob F. Fast   fast@imes.uni-hannover.de
Institute of Mechatronic Systems, Leibniz Universität Hannover, Hannover, Germany
Department of Phoniatrics and Pediatric Audiology, Hannover Medical School, Hannover, Germany

Lüder A. Kahrs   lakahrs@cs.toronto.edu
Centre for Image Guided Innovation and Therapeutic Intervention, The Hospital for Sick Children, Toronto, Canada
Department of Mathematical and Computational Sciences, University of Toronto Mississauga, Mississauga, Canada

Tobias Ortmaier   ortmaier@imes.uni-hannover.de
Institute of Mechatronic Systems, Leibniz Universität Hannover, Hannover, Germany

Abstract

The consideration of predictive uncertainty in medical imaging with deep learning is of utmost importance. We apply estimation of both aleatoric and epistemic uncertainty by variational Bayesian inference with Monte Carlo dropout to regression tasks and show that predictive uncertainty is systematically underestimated. We apply σ scaling with a single scalar value; a simple, yet effective calibration method for both types of uncertainty. The performance of our approach is evaluated on a variety of common medical regression data sets using different state-of-the-art convolutional network architectures. In our experiments, σ scaling is able to reliably recalibrate predictive uncertainty. It is easy to implement and maintains the accuracy. Well-calibrated uncertainty in regression allows robust rejection of unreliable predictions or detection of out-of-distribution samples. Our source code is available at: github.com/mlaves/well-calibrated-regression-uncertainty

Keywords: Bayesian approximation, variational inference

1. Introduction

Predictive uncertainty should be considered in any medical imaging task that is approached with deep learning. Well-calibrated uncertainty is of great importance for decision-making and is anticipated to increase patient safety. It allows to robustly reject unreliable predictions or out-of-distribution samples. In this paper, we address the problem of miscalibration of regression uncertainty with application to medical image analysis.

For the task of regression, we aim to estimate a continuous target value $\boldsymbol{y} \in \mathbb{R}^d$ given an input image $\boldsymbol{x}$. Regression in medical imaging with deep learning has been applied to forensic age estimation from hand CT/MRI (Halabi et al., 2019; Štern et al., 2016), natural landmark localization (Payer et al., 2019), cell detection in histology (Xie et al., 2018), or instrument pose estimation (Gessert et al., 2018). By predicting the coordinates of object
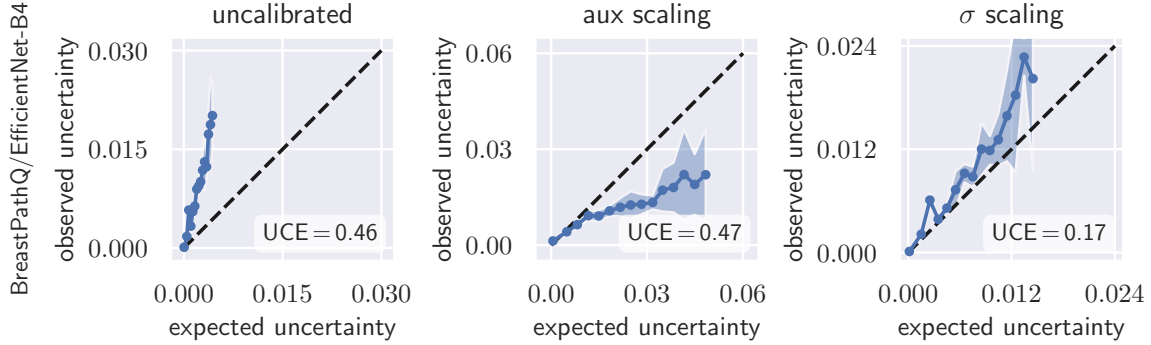
Figure 1: Calibration plots and uncertainty calibration error (UCE) for EfficientNet-B4 on BreastPathQ test set. Uncalibrated uncertainty is underestimated and does not correspond well with the model error (left). Uncertainty can be calibrated most effectively with σ scaling (right). Solid lines show the mean and shaded areas show standard deviation from 5 repeated runs. Dashed lines denote perfect calibration.

boundaries, segmentation can also be performed as a regression task. This has been done for segmentation of pulmonary nodules in CT (Messay et al., 2015), kidneys in ultrasound (Yin et al., 2020), or left ventricles in MRI (Tan et al., 2017). In registration of medical images, a continuous displacement field is predicted for each coordinate of $\boldsymbol{x}$, which has also recently been addressed by CNNs for regression (Dalca et al., 2019).

In medical imaging, it is crucial to consider the predictive uncertainty of deep learning models. Bayesian neural networks (BNN) and their approximation provide mathematical tools for reasoning the uncertainty (Bishop, 2006; Kingma and Welling, 2014). In general, predictive uncertainty can be split into two types: aleatoric and epistemic uncertainty (Tanno et al., 2017; Kendall and Gal, 2017). This distinction was first made in the context of risk management (Hora, 1996). Aleatoric uncertainty arises from the data directly; e.g. sensor noise or motion artifacts. Epistemic uncertainty is caused by uncertainty in the model parameters due to a limited amount of training data (Bishop, 2006). A well-accepted approach to quantify epistemic uncertainty is variational inference with Monte Carlo (MC) dropout, where dropout is used at test time to sample from the approximate posterior (Gal and Ghahramani, 2016).

Uncertainty quantification in regression problems in medical imaging has been addressed by prior work. Medical image enhancement with image quality transfer (IQT) has been extended to a Bayesian approach to obtain pixel-wise uncertainty (Tanno et al., 2016). Additionally, CNN-based IQT was used to estimate both aleatoric and epistemic uncertainty in MRI super-resolution (Tanno et al., 2017). Dalca et al. (2019) estimated uncertainty for a deformation field in medical image registration using a probabilistic CNN. Registration uncertainty has also been addressed outside the deep learning community (Luo et al., 2019). Schlemper et al. (2018) used sub-network ensembles to obtain uncertainty estimates in cardiac MRI reconstruction. Aleatoric and epistemic uncertainty was also used in multitask learning for MRI-based radiotherapy planning (Bragman et al., 2018).

Uncertainty obtained by deep BNNs tends to be miscalibrated, i.e. it does not correlate well with the model error (Laves et al., 2019). Fig. 1 shows calibration plots (observed uncertainty vs. expected uncertainty) for uncalibrated and calibrated uncertainty estimates. The predicted uncertainty (taking into account both epistemic and aleatoric uncertainty) is underestimated and does not allow robust detection of uncertain predictions at test time.

Calibration of uncertainty in regression has been addressed in prior work outside medical imaging. In (Kuleshov et al., 2018), inaccurate uncertainties from Bayesian models for regression are recalibrated using a technique inspired by Platt scaling. Given a pre-trained, miscalibrated model $\boldsymbol{H}$, an auxiliary model $\boldsymbol{R} : [0, 1]^d \to [0, 1]^d$ is trained, that yields a calibrated regressor $\boldsymbol{R} \circ \boldsymbol{H}$. In (Phan et al., 2018), this method was applied to bounding box regression. However, an auxiliary model with enough capacity will always be able to recalibrate, even if the predicted uncertainty is completely uncorrelated with the real uncertainty. Furthermore, Kuleshov et al. (2018) state that calibration via $\boldsymbol{R}$ is possible if enough independent and identically distributed (i.i.d.) data is available. In medical imaging, large data sets are usually hard to obtain, which can cause $\boldsymbol{R}$ to overfit the calibration set. This downside was addressed in (Levi et al., 2019), which is most related to our work. They proposed to scale the standard deviation of a Gaussian model to recalibrate aleatoric uncertainty. In contrast to our work, they do not take into account epistemic uncertainty, which is an important source of uncertainty, especially when dealing with small data sets in medical imaging.

This paper extends a preliminary version of this work presented at the Medical Imaging with Deep Learning (MIDL) 2020 conference (Laves et al., 2020). We continue this work by providing a new derivation of our definition of perfect calibrtaion, new experimental results, analysis and discussion. Additionally, prediction intervals are computed to further assess the quality of the estimated uncertainty. We find that prediction intervals are estimated too narrow and that recalibration can mitigate this problem.

To the best of our knowledge, calibration of predictive uncertainty for regression tasks in medical imaging has not been addressed. Our main contributions are: (1) We suggest to use σ scaling in a separate calibration phase to tackle underestimation of aleatoric and epistemic uncertainty, (2) we propose to use the uncertainty calibration error and prediction intervals to assess the quality of the estimated uncertainty, and (3) we perform extensive experiments on four different data sets to show the effectiveness of the proposed method.

2. Methods

In this section, we discuss estimation of aleatoric and epistemic uncertainty for regression and show why uncertainty is systematically miscalibrated. We propose to use σ scaling to jointly calibrate aleatoric and epistemic uncertainty.

2.1 Conditional Log-Likelihood for Regression

We revisit regression under the maximum posterior (MAP) framework to derive direct estimation of heteroscedastic aleatoric uncertainty. That is, the aleatoric uncertainty varies with the input and is not assumed to be constant. The goal of our regression model is to predict a target value $\boldsymbol{y}$ given some new input $\boldsymbol{x}$ and a training set $\mathcal{D}$ of m inputs $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ and their corresponding (observed) target values $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$. We

assume that $\boldsymbol{y}$ has a Gaussian distribution $\mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}(\boldsymbol{x}), \hat{\sigma}^2(\boldsymbol{x})\right)$ with mean equal to $\hat{\boldsymbol{y}}(\boldsymbol{x})$ and variance $\hat{\sigma}^2(\boldsymbol{x})$. A neural network with parameters $\boldsymbol{\theta}$

$$\boldsymbol{f_\theta}\left(\boldsymbol{x}\right) = \left[\hat{\boldsymbol{y}}(\boldsymbol{x}), \hat{\sigma}^2(\boldsymbol{x})\right], \ \hat{\boldsymbol{y}} \in \mathbb{R}^d, \ \hat{\sigma}^2 \in \mathbb{R}, \hat{\sigma}^2 \geq 0 \tag{1}$$

outputs these values for a given input (Nix and Weigend, 1994). By assuming a Gaussian prior over the parameters $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{0}, \lambda^{-1}\boldsymbol{I})$, MAP estimation becomes maximum-likelihood estimation with added weight decay (Bishop, 2006). With m i.i.d. random samples, the conditional log-likelihood $\log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})$ is given by

$$\log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}_{\boldsymbol{\theta}}^{(i)}} \exp\left\{-\frac{\left\|\boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)}\right\|^2}{2\left(\hat{\sigma}_{\boldsymbol{\theta}}^{(i)}\right)^2}\right\}\right) \tag{2}$$

$$= -\frac{m}{2}\log\left(2\pi\right) - \sum_{i=1}^{m} \log\left(\hat{\sigma}_{\boldsymbol{\theta}}^{(i)}\right) + \frac{1}{2\left(\hat{\sigma}_{\boldsymbol{\theta}}^{(i)}\right)^2}\left\|\boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)}\right\|^2 . \tag{3}$$

The dependence on $\boldsymbol{x}$ has been omitted to simplify the notation. Maximizing the log-likelihood in Eq. (3) w.r.t. $\boldsymbol{\theta}$ is equivalent to minimizing the negative log-likelihood (NLL), which leads to the following optimization criterion (with weight decay)

$$\mathcal{L}_{\mathrm{G}}(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left(\hat{\sigma}_{\boldsymbol{\theta}}^{(i)}\right)^{-2}\left\|\boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)}\right\|^2 + \log\left((\hat{\sigma}_{\boldsymbol{\theta}}^{(i)})^2\right) . \tag{4}$$

Here, $\hat{\boldsymbol{y}}_{\boldsymbol{\theta}}$ and $\hat{\sigma}_{\boldsymbol{\theta}}$ are estimated jointly by finding $\boldsymbol{\theta}$ that minimizes Eq. (4). This can be achieved using gradient descent in a standard training procedure. In this case, $\hat{\sigma}_{\boldsymbol{\theta}}$ captures the uncertainty that is inherent in the data (aleatoric uncertainty). To avoid numerical instability due to potential division by zero, we directly estimate $\log \hat{\sigma}^2(\boldsymbol{x})$ and implement Eq. (4) in similar practice to Kendall and Gal (2017).

2.2 Biased estimation of σ

Ignoring their dependence through $\boldsymbol{\theta}$, the solution to Eq. (4) decouples estimation of $\hat{\boldsymbol{y}}$ and $\hat{\sigma}$. In case of a Gaussian likelihood, minimizing Eq. (4) w.r.t. $\hat{\boldsymbol{y}}^{(i)}$ yields

$$\hat{\boldsymbol{y}}^{(i)} = \underset{\hat{\boldsymbol{y}}^{(i)}}{\arg\min}\, \mathcal{L}_{\mathrm{G}} = \boldsymbol{y}^{(i)} \ \forall i . \tag{5}$$

Minimizing (4) w.r.t. $(\hat{\sigma}^{(i)})^2$ yields

$$\left(\hat{\sigma}^{(i)}\right)^2 = \underset{(\hat{\sigma}^{(i)})^2}{\arg\min}\, \mathcal{L}_{\mathrm{G}} = \|\boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}^{(i)}\|^2 \ \forall i . \tag{6}$$

That is, estimation of σ^2 should perfectly reflect the squared error. However, in Eq. (6) σ^2 is estimated relative to the estimated mean $\hat{\boldsymbol{y}}$ and therefore biased. In fact, the maximum likelihood solution systematically underestimates σ^2, which is a phenomenon of overfitting the training set (Bishop, 2006). The squared error $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$ will be lower on the training set and $\hat{\sigma}^2$ on new samples will be systematically too low (see Fig. 2). This is a problem especially in deep learning, where large models have millions of parameters and tend to overfit. To solve this issue, we introduce a simple learnable scalar parameter s to rescale the biased estimation of σ^2.
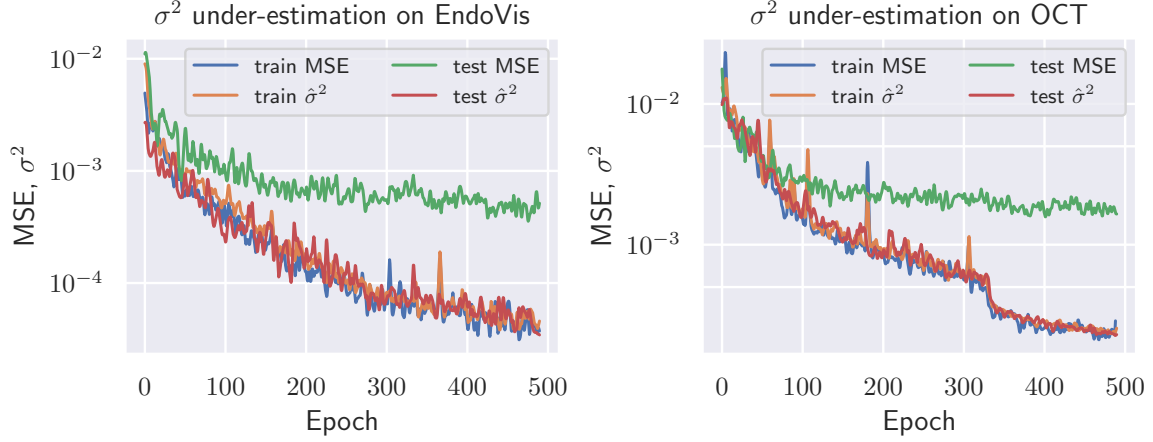
Figure 2: Biased estimation of aleatoric uncertainty σ^2. The deep model overfits estimation of $\boldsymbol{y}$ on the training set. On unseen test data, the MSE of predictive mean is higher and σ^2 is underestimated. Early stopping (e.g. at epoch 50) would result in an unbiased estimator, but this would not be optimal in terms of test MSE.

2.3 σ Scaling for Aleatoric Uncertainty

We first derive σ scaling for aleatoric uncertainty. Using a Gaussian model, we scale the standard deviation σ with a scalar value s to recalibrate the probability density function

$$p\left(\boldsymbol{y}|\boldsymbol{x}; \hat{\boldsymbol{y}}(\boldsymbol{x}), \hat{\sigma}^2(\boldsymbol{x})\right) = \mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}(\boldsymbol{x}), (s \cdot \hat{\sigma}(\boldsymbol{x}))^2\right) \ . \tag{7}$$

This results in the following minimization objective:

$$\mathcal{L}_{\mathrm{G}}(s) = m\log(s) + \tfrac{1}{2}s^{-2}\sum_{i=1}^{m}\left(\hat{\sigma}_{\boldsymbol{\theta}}^{(i)}\right)^{-2}\left\|\boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)}\right\|^2 \ . \tag{8}$$

Eq. (8) is optimized w.r.t. s with fixed $\boldsymbol{\theta}$ using gradient descent in a separate calibration phase after training to calibrate aleatoric uncertainty measured by $\hat{\sigma}_{\boldsymbol{\theta}}^2$. In case of a single scalar, the solution to Eq. (8) can also be written in closed form as

$$s = \pm\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{\sigma}_{\boldsymbol{\theta}}^{(i)}\right)^{-2}\left\|\boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)}\right\|^2} \ . \tag{9}$$

We apply σ scaling to jointly calibrate aleatoric and epistemic uncertainty in the next section.

2.4 Well-Calibrated Estimation of Predictive Uncertainty

So far we have assumed a MAP point estimate for $\boldsymbol{\theta}$ which does not consider uncertainty in the parameters. To quantify both aleatoric and epistemic uncertainty, we extend $\boldsymbol{f_\theta}$

into a fully Bayesian model under the variational inference framework with Monte Carlo dropout (Gal and Ghahramani, 2016). In MC dropout, the model $\boldsymbol{f_{\hat{\theta}}}$ is trained with dropout (Srivastava et al., 2014) and dropout is applied at test time by performing N stochastic forward passes to sample from the approximate Bayesian posterior $\tilde{\boldsymbol{\theta}} \sim q(\boldsymbol{\theta})$. Following (Kendall and Gal, 2017), we use MC integration to approximate the predictive variance

$$\hat{\Sigma}^2 = \underbrace{\frac{1}{N} \sum_{n=1}^{N} \left(\hat{\boldsymbol{y}}_n - \frac{1}{N} \sum_{n=1}^{N} \hat{\boldsymbol{y}}_n \right)^2}_{\text{epistemic}} + \underbrace{\frac{1}{N} \sum_{n=1}^{N} \hat{\sigma}_n^2}_{\text{aleatoric}} \tag{10}$$

and use $\hat{\Sigma}^2$ as a measure of predictive uncertainty. If the neural network has multiple outputs ($d > 1$), the predictive variance is calculated per output and the mean across d forms the final uncertainty value. Eq. (10) is an unbiased estimator of the approximate predictive variance (see proof in Appendix C). From Eq. (33) of our proof follows, that $\hat{\Sigma}^2$ is expected to equal the true variance $\Sigma = \mathbb{E}[(\hat{\boldsymbol{y}} - \boldsymbol{y})^2]$. Thus, we define perfect calibration of regression uncertainty as

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left[\mathbb{E}[(\hat{\boldsymbol{y}} - \boldsymbol{y})^2] \,\middle|\, \hat{\Sigma}^2 = \alpha^2 \right] = \alpha^2 \quad \forall \left\{ \alpha^2 \in \mathbb{R} \,\middle|\, \alpha^2 \geq 0 \right\} \,, \tag{11}$$

which extends the definition of (Levi et al., 2019) to both aleatoric and epistemic uncertainty. We expect that additionally accounting for epistemic uncertainty is particularly beneficial for smaller data sets. However, even in deep learning with Bayesian principles, the approximate posterior predictive distribution can overfit on small data sets. In practice, this leads to underestimation of the predictive uncertainty.

One could regularize overfitting by early stopping that prevents large differences between training and test loss, which would circumvent underestimation of σ^2. However, our experiments show that early stopping is not optimal with regard to accuracy, i.e. the squared error of $\hat{\boldsymbol{y}}$ on both training and testing data (see Fig. 2). In contrast, the model with lowest mean error on the validation set underestimates predictive uncertainty considerably. Therefore, we apply σ scaling to recalibrate the predictive uncertainty $\hat{\Sigma}^2$. This allows a lower squared error while reducing underestimation of uncertainty as shown experimentally in the following section.

2.5 Expected Uncertainty Calibration Error for Regression

We extend the definition of miscalibrated uncertainty for classification (Laves et al., 2019) to quantify miscalibration of uncertainty in regression

$$\mathbb{E}_{\hat{\Sigma}^2} \left[\left| \left(\mathbb{E}[(\hat{\boldsymbol{y}} - \boldsymbol{y})^2] \,\middle|\, \hat{\Sigma}^2 = \alpha^2 \right) - \alpha^2 \right| \right] \quad \forall \left\{ \alpha^2 \in \mathbb{R} \,\middle|\, \alpha^2 \geq 0 \right\} \,, \tag{12}$$

using the second moment of the error. On finite data sets, this can be approximated with the expected uncertainty calibration error (UCE) for regression. Following (Guo et al., 2017), the uncertainty output $\hat{\Sigma}^2$ of a deep model is partitioned into K bins with equal width. A weighted average of the difference between the variance and predictive uncertainty is used:

$$\text{UCE} := \sum_{k=1}^{K} \frac{|B_k|}{m} \left| \text{var}(B_k) - \text{uncert}(B_k) \right| \,, \tag{13}$$

with number of inputs m and set of indices B_k of inputs, for which the uncertainty falls into the bin k. The variance per bin is defined as

$$\mathrm{var}(B_k) := \frac{1}{|B_k|} \sum_{i \in B_m} \frac{1}{N} \sum_{n=1}^{N} (\hat{\boldsymbol{y}}_{i,n} - \boldsymbol{y}_i)^2 \, , \tag{14}$$

with N stochastic forward passes, and the uncertainty per bin is defined as

$$\mathrm{uncert}(B_k) := \frac{1}{|B_k|} \sum_{i \in B_k} \hat{\Sigma}_i^2 \, . \tag{15}$$

Note that computing the second moment from Eq. (12) also incorporates MC samples, which can introduce some bias in the evaluation. The UCE considers both aleatoric and epistemic uncertainty and is given in % throughout this work. Additionally, we plot $\mathrm{var}(B_k)$ vs. $\mathrm{uncert}(B_k)$ to create calibration diagrams.

3. Experiments

We use four data sets and three common deep network architectures to evaluate recalibration with σ scaling. The data sets were selected to represent various regression tasks in medical imaging with different dimension d of target value $\boldsymbol{y} \in \mathbb{R}^d$:

(1) Estimation of tumor cellularity in histology whole slides of cancerous breast tissue from the BreastPathQ SPIE challenge data set ($d = 1$) (Martel et al., 2019). The public data set consists of 2579 images, from which 1379/600/600 are used for training/validation/testing. The ground truth label is a single scalar $y \in [0, 1]$ denoting the ratio of tumor cells to non-tumor cells.

(2) Hand CT age regression from the RSNA pediatric bone age data set ($d = 1$) (Halabi et al., 2019). The task is to infer a person's age in months from CT scans of the hand. This data set is the largest used in this paper and has 12,811 images, from which we use 6811/2000/4000 images for training/validation/testing.

(3) Surgical instrument tracking on endoscopic images from the EndoVis endoscopic vision challenge 2015[1] data set ($d = 2$). This data set contains 8,984 video frames from 6 different robot-assisted laparoscopic interventions showing surgical instruments with ground truth pixel coordinates of the instrument's center point $\boldsymbol{y} \in \mathbb{R}^2$. We use 4483/2248/2253 frames for training/validation/testing. As the public data set is only sparsely annotated, we created our own ground truth labels, which can be found in our code repository.

(4) 6DoF needle pose estimation on optical coherence tomography (OCT) scans from our own data set[2]. This data set contains 5,000 3D-OCT scans with the accompanying needle pose $\boldsymbol{y} \in \mathbb{R}^6$, from which we use 3300/850/850 for training/validation/testing. Additional details on this data set can be found in Appendix E.

All outputs are normalized such that $\boldsymbol{y} \in [0, 1]^d$. The employed network architectures are ResNet-101, DenseNet-201 and EfficientNet-B4 (He et al., 2016; Huang et al., 2017; Tan and Le, 2019). The last linear layer of all networks is replaced by two linear layers

---

1. endovissub-instrument.grand-challenge.org
2. Our OCT pose estimation data set is publicly available at github.com/mlaves/3doct-pose-dataset
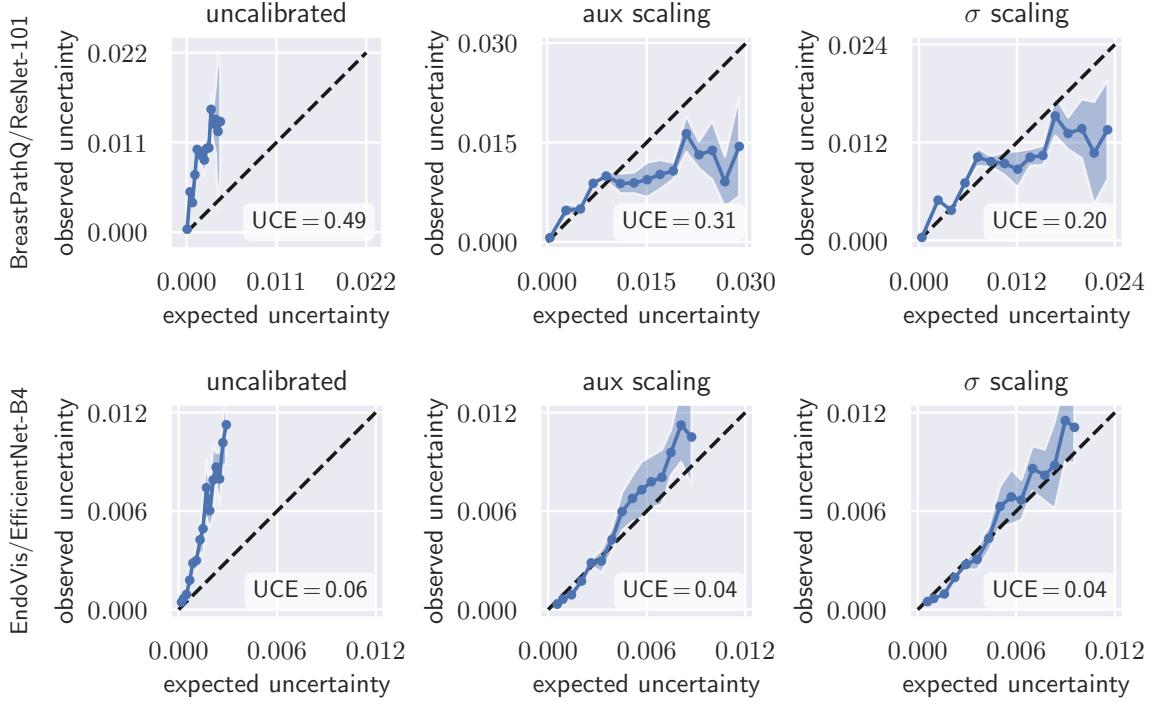
Figure 3: Calibration plots for ResNet-101 on BreastPathQ (top row) and EfficientNet-B4 on EndoVis (bottom row). Aux scaling tends to overfit the calibration set, which results in higher UCE compared to simple σ scaling. Dashed lines denote perfect calibration.

predicting $\hat{\boldsymbol{y}}$ and $\hat{\sigma}^2$ as described in § 2.1. For MC dropout, we use dropout before the last linear layers. Dropout is further added after each of the four layers of stacked residual blocks in ResNet. In DenseNet and EfficientNet, we use the default configuration of dropout during training and testing. The networks are trained until no further decrease in mean squared error (MSE) on the validation set can be observed. More details on the training procedure can be found in Appendix D.

Calibration is performed after training in a separate calibration phase using the validation data set. We plug the predictive uncertainty $\hat{\Sigma}^2$ into Eq. (8) and minimize w.r.t. s. Additionally, we compare σ scaling to a more powerful auxiliary recalibration model $\boldsymbol{R}$ consisting of a two-layer fully-connected network with 16 hidden units and ReLU activations (inspired by (Kuleshov et al., 2018), see § 1).

4. Results

To quantify miscalibration, we use the proposed expected uncertainty calibration error for regression. We visualize (mis-)calibration in Fig. 1 and Fig. 3 using calibration diagrams, which show expected uncertainty vs. observed uncertainty. The discrepancy to the identity function reveals miscalibration. The calibration diagrams clearly show the underestima-

tion of uncertainty for the uncalibrated models. After calibration with both aux and σ scaling, the estimated uncertainty better reflects the actual uncertainty. Figures for all configurations are listed in Appendix G.

Tab. 1 reports UCE values of all data set/model combinations on the respective test sets. The negative log-likelihood also measures miscalibration; the values on the test set can be found in Tab. 2 in the appendix. In general, recalibration considerably reduces miscalibration. On the data sets BoneAge, EndoVis and OCT, both scaling methods perform similarly well. However, on the BreastPathQ data set, σ scaling clearly outperforms aux scaling in terms of UCE. BreastPathQ is the smallest data set and thus has the smallest calibration set size. We hypothesize that the more powerful auxiliary model $\boldsymbol{R}$ overfits the calibration set (see BreastPathQ/DenseNet-201 in Tab. 1), which leads to an increase of UCE on the test set. An ablation study on BreastPathQ for the auxiliary model can be found in Appendix F.

We also compare our approach to Levi et al. (2019) in Tab. 1, which only considers aleatoric uncertainty. The aleatoric uncertainty is well-calibrated if it reflects the bias $(\mathbb{E}\left[\hat{\boldsymbol{y}}_n\right] - \boldsymbol{y})^2$, which is given by the squared error between the expectation of the stochastic predictions $\hat{\boldsymbol{y}}_n$ and the ground truth. Therefore, the UCE for aleatoric-only is computed by $\text{UCE} = \sum_{k=1}^{K} \frac{|B_k|}{m} \left| \text{err}(B_k) - \text{uncert}(B_k) \right|$, where $\text{err}(\cdot)$ is the mean squared error and $\text{uncert}(\cdot)$ is the mean aleatoric uncertainty per bin. Consideration of epistemic uncertainty is especially beneficial on smaller data sets (BreastPathQ), where our approach outperforms Levi et al. (2019). On larger data sets, the benefit diminishes and both approaches are equally calibrated.

Additionally, we report UCE values from a DenseNet ensemble for comparison. In contrast to what is reported by Lakshminarayanan et al. (2017), the deep ensemble tends to be calibrated worse. Only on BoneAge, the ensemble is better calibrated prior to recalibration of the other methods. After recalibration, both approaches outperform the deep ensemble.

Fig. 4 shows the result of intra-training calibration of aleatoric uncertainty. It indicates that the gap between training and test loss is successfully closed.

4.1 Posterior Prediction Intervals

In addition to the calibration diagrams, we compute prediction intervals from the uncalibrated and calibrated posterior predictive distribution. Well-calibrated prediction intervals provide a reliable measure of precision of the estimated target value. In Bayesian inference, prediction intervals define an interval within which the true target value $\boldsymbol{y}^*$ of a new, unobserved input $\boldsymbol{x}^*$ is expected to fall with a specific probability (Heskes, 1997; Held and Sabanés Bové, 2014). This is also referred to as the credible interval of the posterior predictive distribution. For $\gamma \in (0, 1)$, a $\gamma \cdot 100\,\%$ prediction interval is defined through z_l and z_u such that

$$\int_{z_l}^{z_u} p(\boldsymbol{y}^* \,|\, \boldsymbol{x}^*, \mathcal{D})\, \mathrm{d}\boldsymbol{y}^* = \gamma \;, \tag{16}$$

with posterior predictive distribution $p(\boldsymbol{y}^* \,|\, \boldsymbol{x}^*, \mathcal{D})$. We compute the $50\,\%$, $90\,\%$, $95\,\%$, and $99\,\%$ prediction interval using the root of the predictive variance from Eq. (10); that is, the $\hat{\boldsymbol{y}} \pm z\hat{\Sigma}$ intervals with $z \in \{\Phi(0.5), \Phi(0.9), \Phi(0.95), \Phi(0.99)\}$ (estimated interval), with probit function $\Phi(p) = \sqrt{2}\text{erf}^{-1}(p)$ and $\text{erf}(p)$ is the Gaussian error function. This assumes
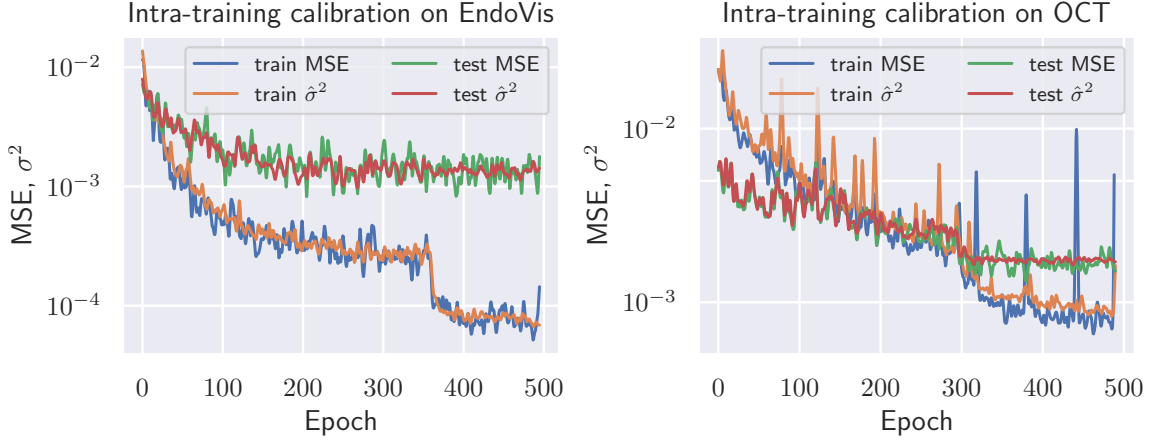
Figure 4: (Left) Intra-training calibration of aleatoric uncertainty with σ scaling. The deep model no longer underestimates $\hat{\sigma}^2$ on unseen test data. (Right) The MSE of predictive mean is higher and σ^2 is underestimated. Note: Calibration is only applied at test time.
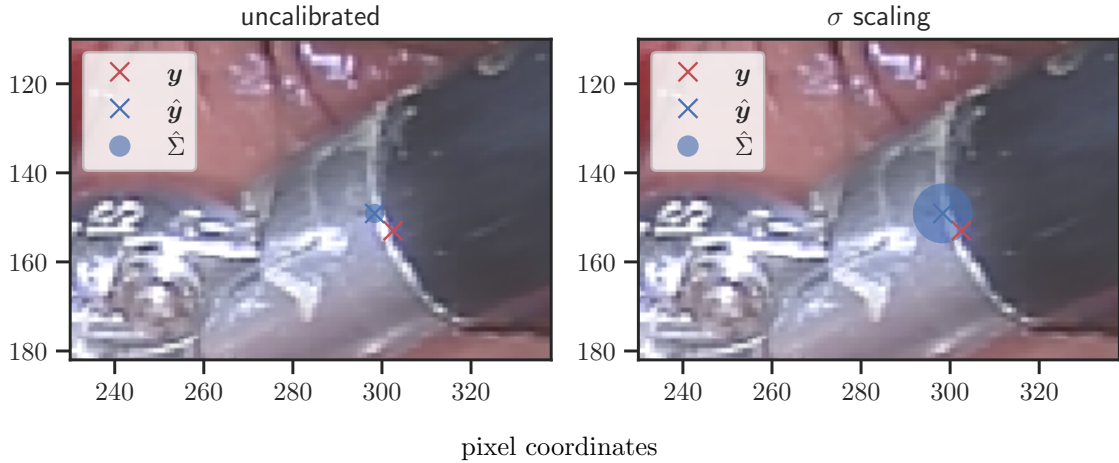

Figure 5: Example result from EndoVis test set. The task is to predict pixel coordinates of the forceps shaft center. Before calibration, the uncertainty is underestimated and the true instrument position $\boldsymbol{y}$ does not fall into the prediction region $\hat{\boldsymbol{y}} \pm \hat{\Sigma}$. After calibration with σ scaling, the uncertainty better reflects the predictive error.

Table 1: Uncertainty calibration error test set results for different datasets and model architectures (averaged over 5 runs). High UCE values indicate miscalibration. In addition, the resulting s for σ scaling is given. We also report UCE values for an ensemble of DenseNets. Bold font indicates lowest values in each experiment.

Data Set	Model	MSE	Levi et al.				ours				ensemble
			none	aux	σ	s	none	aux	σ	s	
BreastPathQ	ResNet-101	6.4e-3	0.51	0.35	0.28	2.91	0.49	0.31	**0.20**	2.37	0.51
	DenseNet-201	7.0e-3	0.21	0.38	**0.15**	1.62	**0.11**	0.36	0.15	1.33	
	EfficientNet-B4	6.4e-3	0.49	0.65	**0.10**	2.30	0.46	0.47	0.17	1.77	
BoneAge	ResNet-101	5.3e-3	0.28	0.07	0.06	1.46	0.28	**0.02**	0.06	1.40	0.09
	DenseNet-201	3.5e-3	0.31	**0.05**	**0.05**	2.98	0.31	**0.05**	**0.05**	2.54	
	EfficientNet-B4	3.5e-3	0.30	0.05	0.10	4.83	0.30	**0.03**	0.12	3.98	
EndoVis	ResNet-101	4.0e-4	**0.04**	0.10	0.09	6.07	**0.04**	**0.04**	**0.04**	3.50	0.08
	DenseNet-201	1.1e-3	0.09	0.05	0.05	3.24	**0.04**	**0.04**	**0.04**	2.57	
	EfficientNet-B4	8.9e-4	0.06	0.05	0.06	2.25	0.06	**0.04**	**0.04**	1.79	
OCT	ResNet-101	2.0e-3	0.17	0.02	0.02	2.74	0.17	**0.01**	0.02	2.14	0.67
	DenseNet-201	1.3e-3	0.08	**0.01**	0.02	1.60	0.04	0.03	0.02	1.26	
	EfficientNet-B4	1.4e-3	0.12	**0.01**	**0.01**	2.65	0.12	**0.01**	**0.01**	1.94	

that the posterior predictive distribution is Gaussian, which is not generally the case. To assess the calibration of the posterior prediction interval, we compute the percentage of how many of the ground truth values of the test set actually fall within the respective intervals (observed interval). In Fig. 6, selected plots of observed vs. estimated prediction intervals are shown. A complete list of prediction intervals can be found in Appendix G.1.

In general, the uncalibrated prediction intervals are estimated to be too narrow, which is a direct consequence of the underestimated predictive variance. For example, the uncalibrated 90 % interval on DenseNet-201/BoneAge actually only contains approx. 50 % of the ground truth values. On this data set, the prediction intervals are considerably improved after recalibration (Fig. 6 left). If a network is already well-calibrated, recalibration can lead to overestimation of the lower prediction intervals (Fig. 6 right). However, in all cases, the 99 % prediction interval contains approx. 99 % of the ground truth test set values after recalibration. This is not the case without the proposed calibration methods. Fig. 5 shows a practical example of the $\hat{y} \pm \hat{\Sigma}$ prediction region from the EndoVis test set. Even though the posterior predictive distribution is not necessarily Gaussian, the calibrated results fit the prediction intervals well. This is especially the case for BoneAge, which is the largest data set used in this paper.

4.2 Detection of Out-of-Distribution Data and Unreliable Predictions

Deep neural networks only yield reliable predictions for data which follow the same distribution as the training data. A shift in distribution could occur when a model trained on CT data from a specific CT device is applied to data from another manufacturer's CT device, for example. This could potentially lead to wrong predictions with low uncertainty, which we tackle with recalibration. To create a moderate distribution shift, we preprocess images from the BoneAge data set using Contrast Limited Adapative Histogram Equalization

(CLAHE) (Pizer et al., 1987) with a clip-limit of 0.03 and report histograms of the uncertainties (see Fig. 7). Additionally, a severe distribution shift is created by presenting images from the BreastPathQ data set to the models trained on BoneAge. Lakshminarayanan et al. (2017) state that deep ensembles provide better-calibrated uncertainty than Bayesian neural networks with MC dropout variational inference. We therefore train an ensemble of 5 randomly initialized DenseNet-201 and compare Bayesian uncertainty with σ scaling to ensemble uncertainty under distribution shift. The results with σ scaling are comparable to those from a deep ensemble for a moderate shift, but without the need to train multiple models on the same data set. A severe shift leads to only slightly increased uncertainties from the calibrated MC dropout model, while the deep ensemble is more sensitive.

Additionally, we apply the well-calibrated models to detect and reject uncertain predictions, as crucial decisions in medical practice should only be made on the basis of reliable predictions. An uncertainty threshold Σ_{max}^2 is defined and all predictions from the test set are rejected where $\hat{\Sigma}^2 > \Sigma_{\mathrm{max}}^2$ (see Fig. 8). From this, a decrease in overall MSE is expected. We additionally compare rejection on the basis of σ scaled uncertainty to uncertainty from the aforementioned ensemble. In case of σ scaling, the test set MSE decreases monotonically as a function of the uncertainty threshold, whereas the ensemble initially shows an increasing MSE (see Fig. 8).

5. Discussion & Conclusion

In this paper, well-calibrated predictive uncertainty in medical imaging obtained by variational inference with deep Bayesian models is discussed. Both aux and σ scaling calibration methods considerably reduce miscalibration of predictive uncertainty in terms of UCE. If the deep model is already well-calibrated, σ scaling does not negatively affect the calibration, which results in $s \to 1$. More complex calibration methods such as aux scaling have to be used with caution, as they can overfit the data set used for calibration. If the calibration set is sufficiently large, they can outperform simple scaling. However, models trained on large data sets are generally better calibrated and the benefit diminishes. Compared to the work of Levi et al. (2019), accounting for epistemic uncertainty is particularly beneficial for smaller data sets, which is helpful in medical practice where access to large labeled data sets is less common and is associated with great costs.

Posterior prediction intervals provide another insight into the calibration of deep models. After recalibration, the 99 % posterior prediction intervals correctly contain approx. 99 % of the ground truth test set values. In some cases, lower prediction intervals are estimated to be too wide after calibration. This is especially the case for smaller data sets and we conjecture that small calibration sets may not contain enough i.i.d. data for calibrating lower prediction intervals and that the assumption of a Gaussian predictive distribution is too strong in this case. On the smallest data set BreastPathQ, aux scaling seems to perform better in terms of prediction intervals, but not in terms of UCE.

Well-calibrated uncertainties from MC dropout are able to detect a moderate shift in the data distribution. However, deep ensembles perform better under a severe distribution shift. BNNs with calibrated uncertainty by σ scaling outperform ensemble uncertainty in the rejection task, which we attribute to the generally poorer calibration of ensembles on in-distribution data.
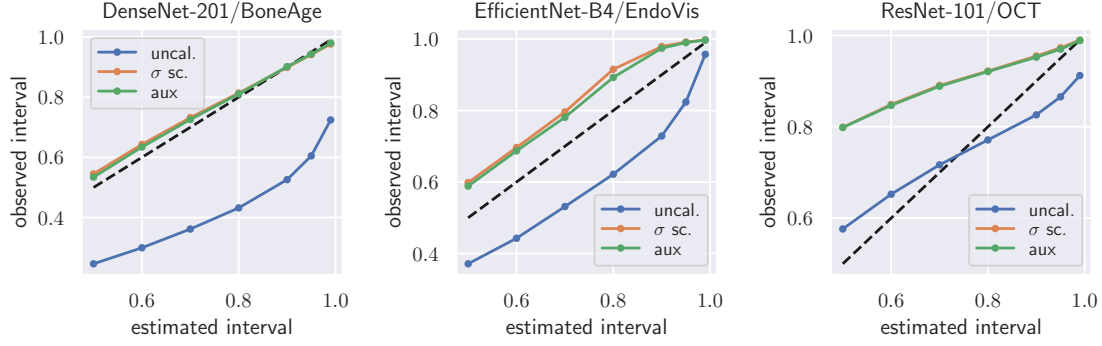
Figure 6: Observed vs. estimated posterior prediction intervals on the test sets. (Left & center) The uncalibrated prediction interval is too narrow due to underestimation of uncertainty. (Right) Calibration can lead to overestimation of predictive intervals, if the network is already well-calibrated. Dashed lines denote 1:1 mapping.
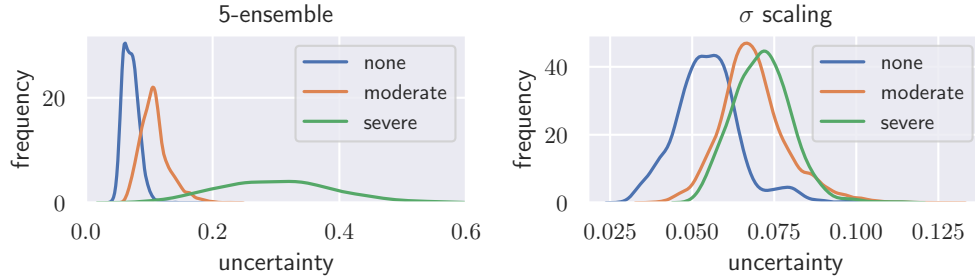

Figure 7: Histograms of the uncertainties for out-of-distribution detection with DenseNet-201 on BoneAge test set. (Left) Uncertainties from a non-Bayesian ensemble of five DenseNets and (right) Bayesian uncertainties calibrated with σ scaling. The distribution shifts have been created with pre-processing by CLAHE (moderate) and images from a different domain (severe).
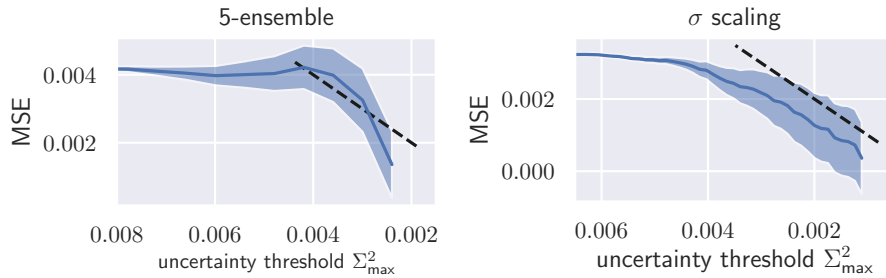

Figure 8: Rejection of uncertain predictions with DenseNet-201 on BoneAge test set with $\hat{\Sigma}^2 > \Sigma^2_{\max}$. The shaded area width visualizes the percentage of rejected samples. The dashed line visualizes linear relationship.

σ scaling is simple to implement, does not change the predictive mean $\hat{\boldsymbol{y}}$, and therefore guarantees to conserve the model's accuracy. It is preferable to regularization (e.g. early stopping) or more complex recalibration methods in calibrated uncertainty estimation with Bayesian deep learning. The disconnection between training and test NLL can successfully be closed, which creates highly accurate models with reliable uncertainty estimates. However, there are many factors (e.g. network capacity, weight decay, dropout configuration) influencing the uncertainty that have not been discussed here and will be addressed in future work.

Acknowledgments

References

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN 978-0-387-31073-2.

Felix JS Bragman, Ryutaro Tanno, Zach Eaton-Rosen, Wenqi Li, David J Hawkes, Sebastien Ourselin, Daniel C Alexander, Jamie R McClelland, and M Jorge Cardoso. Uncertainty in multitask learning: joint representations for probabilistic mr-only radiotherapy planning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–11, 2018.

Adrian V. Dalca, Guha Balakrishnan, John Guttag, and Mert R. Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Med Image Anal*, 57:226–236, 2019. doi: https://doi.org/10.1016/j.media.2019.07.006.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016.

Nils Gessert, Matthias Schlüter, and Alexander Schlaefer. A deep learning approach for pose estimation from volumetric OCT data. *Med Image Anal*, 46:162–179, 2018. doi: 10.1016/j.media.2018.03.002.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330, 2017.

Safwan S. Halabi, Luciano M. Prevedello, Jayashree Kalpathy-Cramer, Artem B. Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, Felipe Campos Kitamura, Hans H. Thodberg, Leon Chen, George Shih, Katherine Andriole, Marc D. Kohli, Bradley J. Erickson, and Adam E. Flanders. The RSNA pediatric bone age machine learning challenge. *Radiol*, 290(2):498–503, 2019. doi: 10.1148/radiol.2018180736.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

Leonhard Held and Daniel Sabanés Bové. *Applied Statistical Inference*. Springer, 1 edition, 2014. ISBN 978-3-642-37887-4. doi: 10.1007/978-3-642-37887-4.

Tom Heskes. Practical confidence and prediction intervals. In *NeurIPS*, pages 176–182, 1997.

Stephen C Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54 (2-3):217–223, 1996.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, pages 5574–5584, 2017.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *ICML*, volume 80, pages 2796–2804, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, pages 6402–6413, 2017.

Max-Heinrich Laves, Andreas Schoob, Lüder A. Kahrs, Tom Pfeiffer, Robert Huber, and Tobias Ortmaier. Feature tracking for automated volume of interest stabilization on 4D-OCT images. In *SPIE Medical Imaging*, volume 10135, pages 256–262, 2017. doi: 10.1117/12.2255090.

Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. In *Bayesian Deep Learning Workshop (NeurIPS)*, 2019. arXiv:1909.13550.

Max-Heinrich Laves, Sontje Ihler, Jacob F Fast, Lüder A Kahrs, and Tobias Ortmaier. Well-calibrated regression uncertainty in medical imaging with deep learning. In *Medical Imaging with Deep Learning*, 2020.

Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. In *arXiv*, 2019. arXiv:1905.11659.

Jie Luo, Alireza Sedghi, Karteek Popuri, Dana Cobzas, Miaomiao Zhang, Frank Preiswerk, Matthew Toews, Alexandra Golby, Masashi Sugiyama, William M Wells, et al. On the applicability of registration uncertainty. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 410–419, 2019.

A. L. Martel, S. Nofech-Mozes, S. Salama, S. Akbar, and M. Peikari. Assessment of residual breast cancer cellularity after neoadjuvant chemotherapy using digital pathology [data set]. *The Cancer Imaging Archive*, 2019. doi: 10.7937/TCIA.2019.4YIBTJNO.

Temesguen Messay, Russell C. Hardie, and Timothy R. Tuinstra. Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the lung image database consortium and image database resource initiative dataset. *Med Image Anal*, 22(1):48–62, 2015. doi: 10.1016/j.media.2015.02.002.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of IEEE International Conference on Neural Networks*, volume 1, pages 55–60, 1994.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019.

Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. *Med Image Anal*, 54:207–219, 2019. doi: 10.1016/j.media.2019.03.007.

Buu Phan, Rick Salay, Krzysztof Czarnecki, Vahdat Abdelzad, Taylor Denouden, and Sachin Vernekar. Calibrating uncertainties in object localization task. In *Bayesian Deep Learning Workshop (NeurIPS)*, 2018. arXiv:1811.11210.

Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.

Jo Schlemper, Guang Yang, Pedro Ferreira, Andrew Scott, Laura-Ann McGill, Zohya Khalique, Margarita Gorodezky, Malte Roehl, Jennifer Keegan, Dudley Pennell, et al. Stochastic deep compressive sensing for the reconstruction of diffusion tensor cardiac mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 295–303, 2018.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15: 1929–1958, 2014.

Darko Štern, Christian Payer, Vincent Lepetit, and Martin Urschler. Automated age estimation from hand mri volumes using deep learning. In *MICCAI*, pages 194–202, 2016.

Li Kuo Tan, Yih Miin Liew, Einly Lim, and Robert A. McLaughlin. Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine MR sequences. *Med Image Anal*, 39:78–86, 2017. doi: 10.1016/j.media.2017.04.002.

Mingxing Tan and Quoc V Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

Ryutaro Tanno, Aurobrata Ghosh, Francesco Grussu, Enrico Kaden, Antonio Criminisi, and Daniel C Alexander. Bayesian image quality transfer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 265–273, 2016.

Ryutaro Tanno, Daniel E Worrall, Aurobrata Ghosh, Enrico Kaden, Stamatios N Sotiropoulos, Antonio Criminisi, and Daniel C Alexander. Bayesian image quality transfer with cnns: exploring uncertainty in dmri super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 611–619, 2017.

Yuanpu Xie, Fuyong Xing, Xiaoshuang Shi, Xiangfei Kong, Hai Su, and Lin Yang. Efficient and robust cell detection: A structured regression approach. *Med Image Anal*, 44:245–254, 2018. doi: 10.1016/j.media.2017.07.003.

Shi Yin, Qinmu Peng, Hongming Li, Zhengqiang Zhang, Xinge You, Katherine Fischer, Susan L. Furth, Gregory E. Tasian, and Yong Fan. Automatic kidney segmentation in ultrasound images using subsequent boundary distance regression and pixelwise classification networks. *Med Image Anal*, 60:101602, 2020. doi: 10.1016/j.media.2019.101602.

Appendix A. Laplacian Model

Using $\mathsf{Laplace}(\hat{\boldsymbol{y}}(\boldsymbol{x}), \hat{\sigma}(\boldsymbol{x}))$ as model, the conditional log-likelihood is given by

$$\log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \sum_{i=1}^{m} \log \left(\frac{1}{2\hat{\sigma}_{\boldsymbol{\theta}}^{(i)}} \exp \left\{ -\frac{\left| \boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)} \right|}{\hat{\sigma}_{\boldsymbol{\theta}}^{(i)}} \right\} \right) \ , \tag{17}$$

which results in the following minimization criterion:

$$\mathcal{L}_{\mathrm{L}}(\boldsymbol{\theta}) = \sum_{i=1}^{m} \frac{1}{\hat{\sigma}_{\boldsymbol{\theta}}^{(i)}} \left| \boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)} \right| + \log \left(\hat{\sigma}_{\boldsymbol{\theta}}^{(i)} \right) \ . \tag{18}$$

Using $\mathcal{L}_{\mathrm{L}}(\boldsymbol{\theta})$ instead of $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\theta})$ results in applying an L1 metric on the predictive mean. In some cases, this led to better results. However, we have not conducted extensive experiments with it and leave it to future work.

Appendix B. Derivation of σ Scaling

See §2.3. Using a Gaussian model, we scale the standard deviation σ with a scalar value s to calibrate the probability density function

$$p\left(\boldsymbol{y} | \boldsymbol{x}; \hat{\boldsymbol{y}}(x), \hat{\sigma}^2(x) \right) = \mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}(x), (s \cdot \hat{\sigma}(x))^2 \right) \ . \tag{19}$$

The conditional log-likelihood is given by

$$\log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \sum_{i=1}^{m} \log \left(\frac{1}{\sqrt{2\pi} s \hat{\sigma}_{\boldsymbol{\theta}}^{(i)}} \exp \left(\frac{\left\| \boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)} \right\|^2}{2 \left(s \hat{\sigma}_{\boldsymbol{\theta}}^{(i)} \right)^2} \right) \right) \tag{20}$$

$$= -\frac{m}{2} \log \left(2\pi \right) - \sum_{i=1}^{m} \log \left(s \hat{\sigma}_{\boldsymbol{\theta}}^{(i)} \right) + \frac{1}{2} \left(s \hat{\sigma}_{\boldsymbol{\theta}}^{(i)} \right)^{-2} \cdot \left\| \boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)} \right\|^2 \tag{21}$$

This results in the following optimization objective (ignoring constants):

$$\mathcal{L}_{\mathrm{G}}(s) = m \log(s) + \tfrac{1}{2} s^{-2} \sum_{i=1}^{m} (\hat{\sigma}_{\boldsymbol{\theta}}^{(i)})^{-2} \left\| \boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)} \right\|^2 \ . \tag{22}$$

Using a Laplacian model, the optimization criterion follows as

$$\mathcal{L}_{\mathrm{L}}(s) = m \log(s) + s^{-1} \sum_{i=1}^{m} \frac{1}{\hat{\sigma}_{\boldsymbol{\theta}}^{(i)}} \left| \boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)} \right| \ . \tag{23}$$

Eq. (22) and (23) are optimized w.r.t. s with fixed $\boldsymbol{\theta}$ using gradient descent in a separate calibration phase after training. The solution to Eq. (22) can also be written in closed form as

$$s_{\mathrm{G}} = \pm \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(\hat{\sigma}_{\boldsymbol{\theta}}^{(i)} \right)^{-2} \left\| \boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)} \right\|^2} \tag{24}$$

and the solution to Eq. (23) follows as

$$s_{\mathrm{L}} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\hat{\sigma}_{\boldsymbol{\theta}}^{(i)}} \left| \boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)} \right| \ , \tag{25}$$

respectively. We apply σ scaling to jointly calibrate aleatoric and epistemic uncertainty as described in § 2.4.

Appendix C. Unbiased Estimator of the Approximate Predictive Variance

We show that the expectation of the predictive sample variance from MC dropout, as given in (Kendall and Gal, 2017), equals the true variance of the approximate posterior predictive distribution.

Proposition 1 *Given N MC dropout samples $\boldsymbol{f}_{\boldsymbol{\theta}_n} = [\hat{\boldsymbol{y}}_n, \hat{\sigma}_n^2]$ from our approximate predictive distribution $p(\boldsymbol{y}^*|\boldsymbol{x}^*, \mathcal{D}) = \mathcal{N}(\boldsymbol{y}^*; \boldsymbol{y}, \Sigma^2)$, the predictive sample variance*

$$\hat{\Sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} \left(\hat{\boldsymbol{y}}_n - \frac{1}{N} \sum_{n=1}^{N} \hat{\boldsymbol{y}}_n \right)^2 + \frac{1}{N} \sum_{n=1}^{N} \hat{\sigma}_n^2 \tag{26}$$

is an unbiased estimator of the approximate predictive variance.

Proof

$$\mathbb{E}\left[\hat{\Sigma}^2\right] = \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} \left(\hat{\boldsymbol{y}}_n - \frac{1}{N} \sum_{n=1}^{N} \hat{\boldsymbol{y}}_n \right)^2 + \frac{1}{N} \sum_{n=1}^{N} \hat{\sigma}_n^2 \right] \tag{27}$$

$$= \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} \left(\hat{\boldsymbol{y}}_n - \frac{1}{N} \sum_{n=1}^{N} \hat{\boldsymbol{y}}_n \right)^2 \right] + \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} \hat{\sigma}_n^2 \right] \tag{28}$$

$$\text{with} \quad \frac{1}{N} \sum_{n=1}^{N} \hat{\boldsymbol{y}}_n = \bar{\boldsymbol{y}} \quad \text{follows} \tag{29}$$

$$= \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} (\hat{\boldsymbol{y}}_n - \bar{\boldsymbol{y}})^2 \right] + \hat{\sigma}^2 \tag{30}$$

$$= \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} (\hat{\boldsymbol{y}}_n - \bar{\boldsymbol{y}})^2 + \bar{\boldsymbol{y}}^2 - \bar{\boldsymbol{y}}^2 + \boldsymbol{y}^2 - \boldsymbol{y}^2 + 2\bar{\boldsymbol{y}}\boldsymbol{y} - 2\bar{\boldsymbol{y}}\boldsymbol{y} \right] + \hat{\sigma}^2 \tag{31}$$

$$= \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} (\hat{\boldsymbol{y}}_n - \boldsymbol{y})^2 - (\bar{\boldsymbol{y}} - \boldsymbol{y})^2 \right] + \hat{\sigma}^2 \tag{32}$$

$$= \mathbb{E}\left[(\hat{\boldsymbol{y}} - \boldsymbol{y})^2 \right] - \mathbb{E}\left[(\bar{\boldsymbol{y}} - \boldsymbol{y})^2 \right] + \hat{\sigma}^2 \qquad\qquad = \Sigma^2 - \hat{\sigma}^2 + \hat{\sigma}^2 \tag{33}$$

$$\mathbb{E}\left[\hat{\Sigma}^2\right] = \Sigma^2 \tag{34}$$

Note that the predicted heteroscedastic aleatoric uncertainty $\hat{\sigma}^2$ equals the bias $\mathbb{E}[(\bar{\boldsymbol{y}} - \boldsymbol{y})^2]$ in Eq. (33) when the aleatoric uncertainty is perfectly calibrated, thus $\mathbb{E}[(\bar{\boldsymbol{y}} - \boldsymbol{y})^2] = \hat{\sigma}^2$. ∎

Appendix D. Training Procedure

The model implementations from PyTorch 1.3 (Paszke et al., 2019) are used and trained with the following settings:

- training for 500 epochs with batch size of 16
- Adam optimizer with initial learn rate of $3 \cdot 10^{-4}$ and weight decay with $\lambda = 10^{-7}$
- reduce-on-plateau learn rate scheduler (patience of 20 epochs) with factor of 0.1
- in MC dropout, $N = 25$ forward passes were performed with dropout with $p = 0.5$ used for ResNet (as described in (Gal and Ghahramani, 2016)). In DenseNet ($p = 0.2$) and EfficientNet ($p = 0.4$) standard dropout p of the architecture is used.
- Additional validation and test sets are used if provided by the data sets; otherwise, a train/validation/test split of approx. $50\% / 25\% / 25\%$ is used
- Source code for all experiments is available at github.com/mlaves/well-calibrated-regression-uncertainty
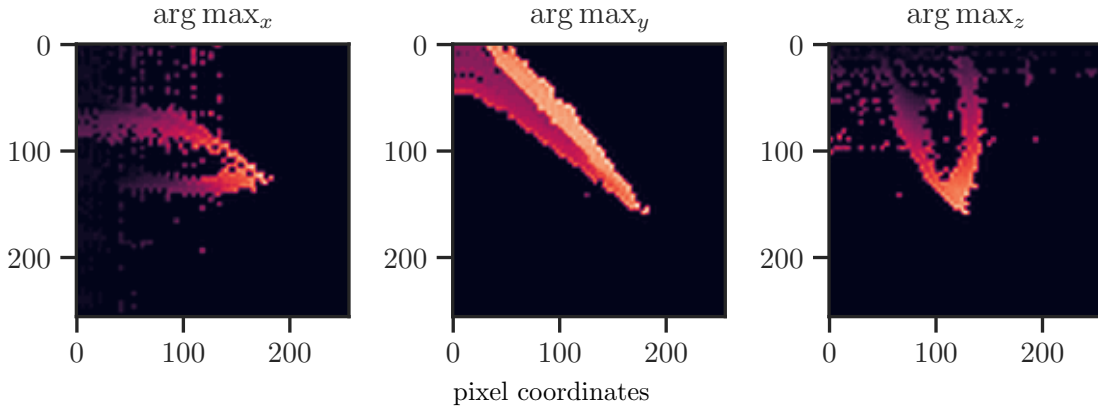
Appendix E. 3D OCT Needle Pose Data Set

Figure 9: Example image from OCT data set showing $\arg\max$ projections of a surgical needle tip acquired by optical coherence tomography.

Our data set was created by attaching a surgical needle to a high-precision six-axis hexapod robot (H-826, Physik Instrumente GmbH & Co. KG, Germany) and observing the needle tip with 3D optical coherence tomography (OCS1300SS, Thorlabs Inc., USA). The data set consists of 5,000 OCT acquisitions with $(64 \times 64 \times 512)$ voxels, covering a volume of approx. $(3 \times 3 \times 3)\,\mathrm{mm}^3$. Each acquisition is taken at a different robot configuration and labeled

with the corresponding 6DoF pose $\boldsymbol{y} \in \mathbb{R}^6$. To process the volumetric data with CNNs for planar images, we calculate 3 planar projections along the spatial dimensions using the $\arg\max$ operator, scale them to equal size and stack them together as three-channel image (see Fig. 9). A similar approach was presented in (Laves et al., 2017) and (Gessert et al., 2018). The data are characterized by a high amount of speckle noise, which is a typical phenomenon in optical coherence tomography. The data set is publicly available at github.com/mlaves/3doct-pose-dataset.

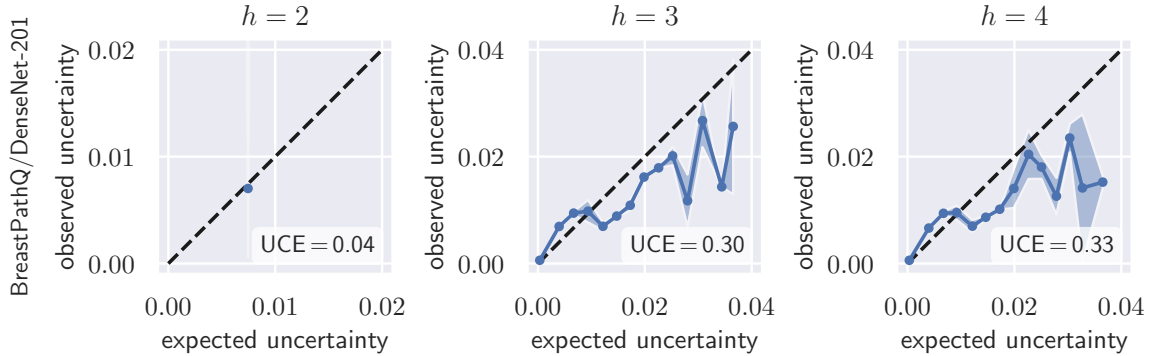## Appendix F. Ablation Study on Auxiliary Model Scaling

Figure 10: Calibration diagrams for aux scaling with different number of hidden layer units h on BreastPathQ/DenseNet-201.

Here, we investigate the overfitting behavior of aux scaling by reducing the number of hidden layer units h of the two-layer auxiliary model with ReLU activations. Aux scaling is more powerful than σ scaling, which can lead to overfitting the calibration set. Fig. 10 shows calibration diagrams for the auxiliary model ablations. Reducing h leads to a minor calibration improvement, but at $h = 2$, the model outputs a constant uncertainty, which is close to the overall mean of the observed uncertainty. A single-layer single-unit model without bias would be equivalent to σ scaling.

## Appendix G. Additional Results and Calibration Diagrams

All test set runs have been repeated 5 times. Solid lines denote mean and shaded areas denote standard deviation calculated from the repeated runs.

Table 2: Negative log-likelihood test set results for different datasets and model architectures (averaged over 5 runs). High NLL values indicate miscalibration. We also report NLL values for an ensemble of DenseNets. Bold font indicates lowest values in each experiment.

| Data Set | Model | MSE | Levi et al. | | | ours | | | ensemble |
			none	aux	σ	none	aux	σ	
BreastPathQ	ResNet-101	6.4e-3	-0.78	-5.06	-5.06	-2.89	**-5.17**	-5.16	
	DenseNet-201	7.0e-3	-5.16	-5.84	-5.70	-5.67	**-6.03**	-5.78	0.11
	EfficientNet-B4	6.4e-3	-3.11	-5.99	-5.53	-4.73	**-6.16**	-5.62	
BoneAge	ResNet-101	5.3e-3	-3.90	**-4.34**	**-4.34**	-3.99	**-4.34**	**-4.34**	
	DenseNet-201	3.5e-3	1.74	**-4.70**	-4.69	-0.75	**-4.70**	-4.69	0.07
	EfficientNet-B4	3.5e-3	13.61	-4.74	-4.67	6.40	**-4.75**	-4.64	
EndoVis	ResNet-101	4.0e-4	-0.53	-6.32	-6.33	-3.85	**-6.76**	-6.72	
	DenseNet-201	1.1e-3	-0.72	**-6.10**	-5.99	-4.94	-6.05	-6.04	0.04
	EfficientNet-B4	8.9e-4	-5.10	-6.06	-6.07	-5.94	**-6.17**	**-6.17**	
OCT	ResNet-101	2.0e-3	-1.08	**-5.24**	**-5.24**	-3.38	**-5.24**	**-5.24**	
	DenseNet-201	1.3e-3	-5.05	-5.61	-5.61	-5.51	**-5.62**	-5.61	0.10
	EfficientNet-B4	1.4e-3	-1.72	**-5.58**	-5.57	-4.25	**-5.58**	-5.57	
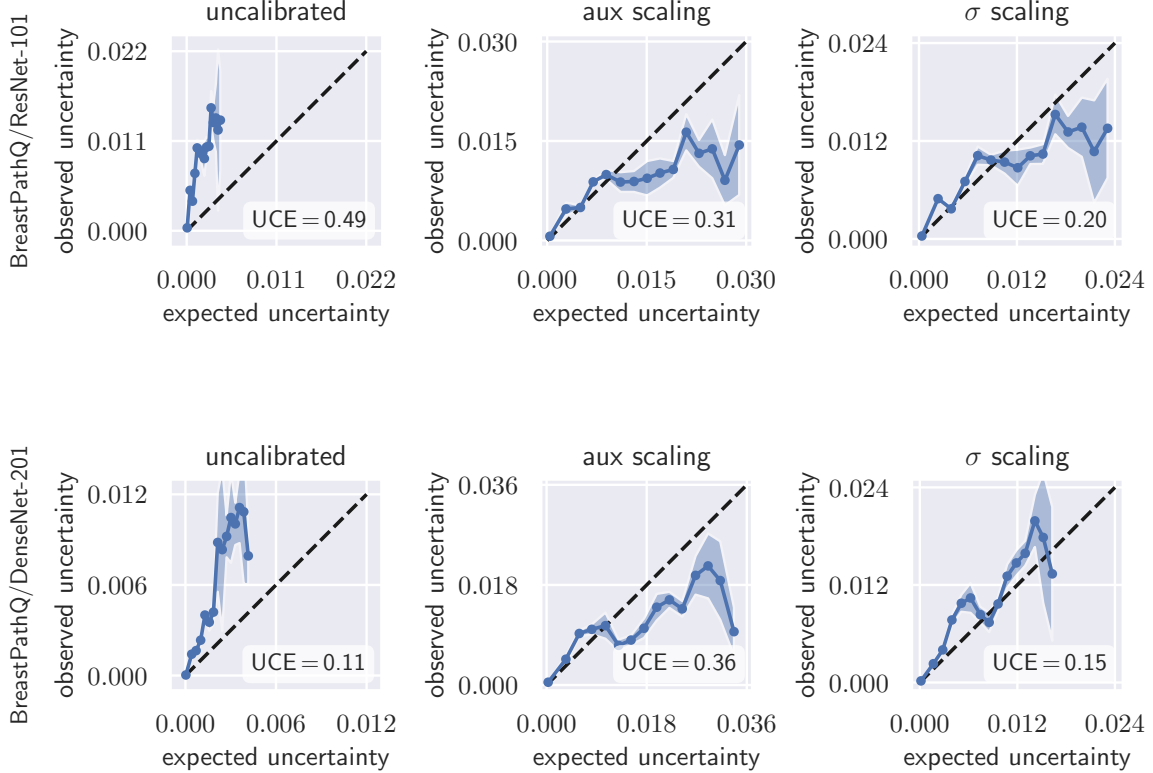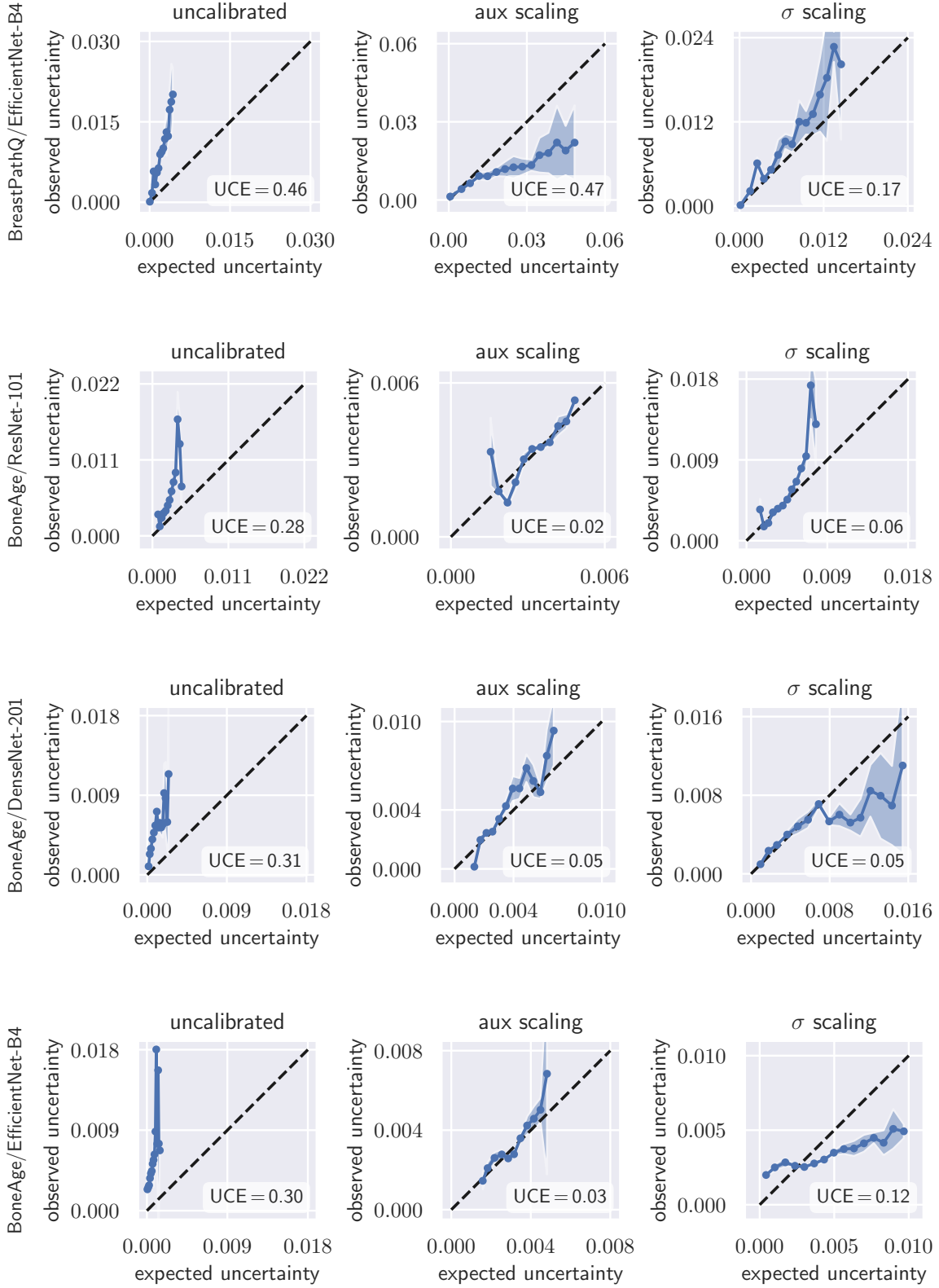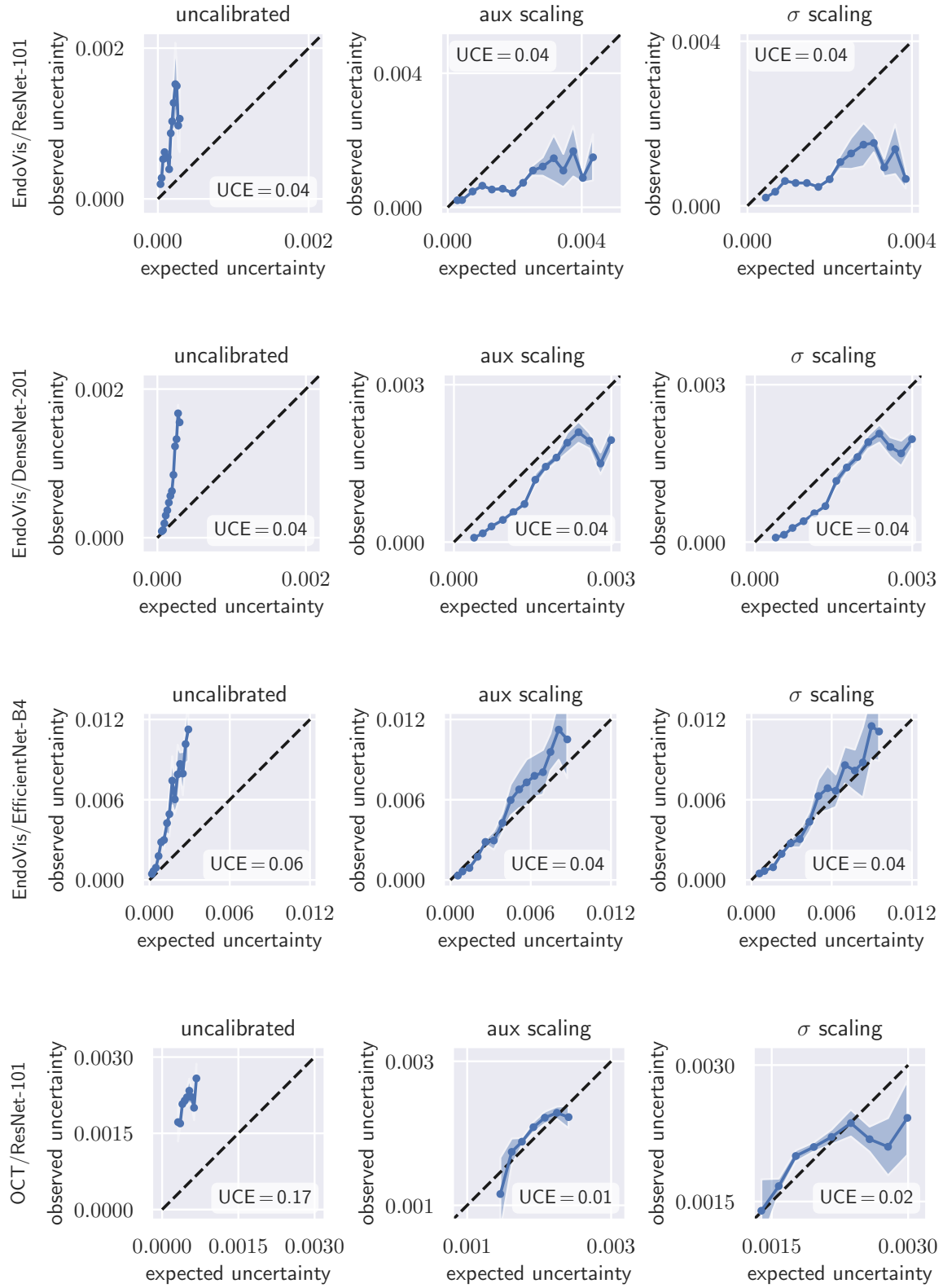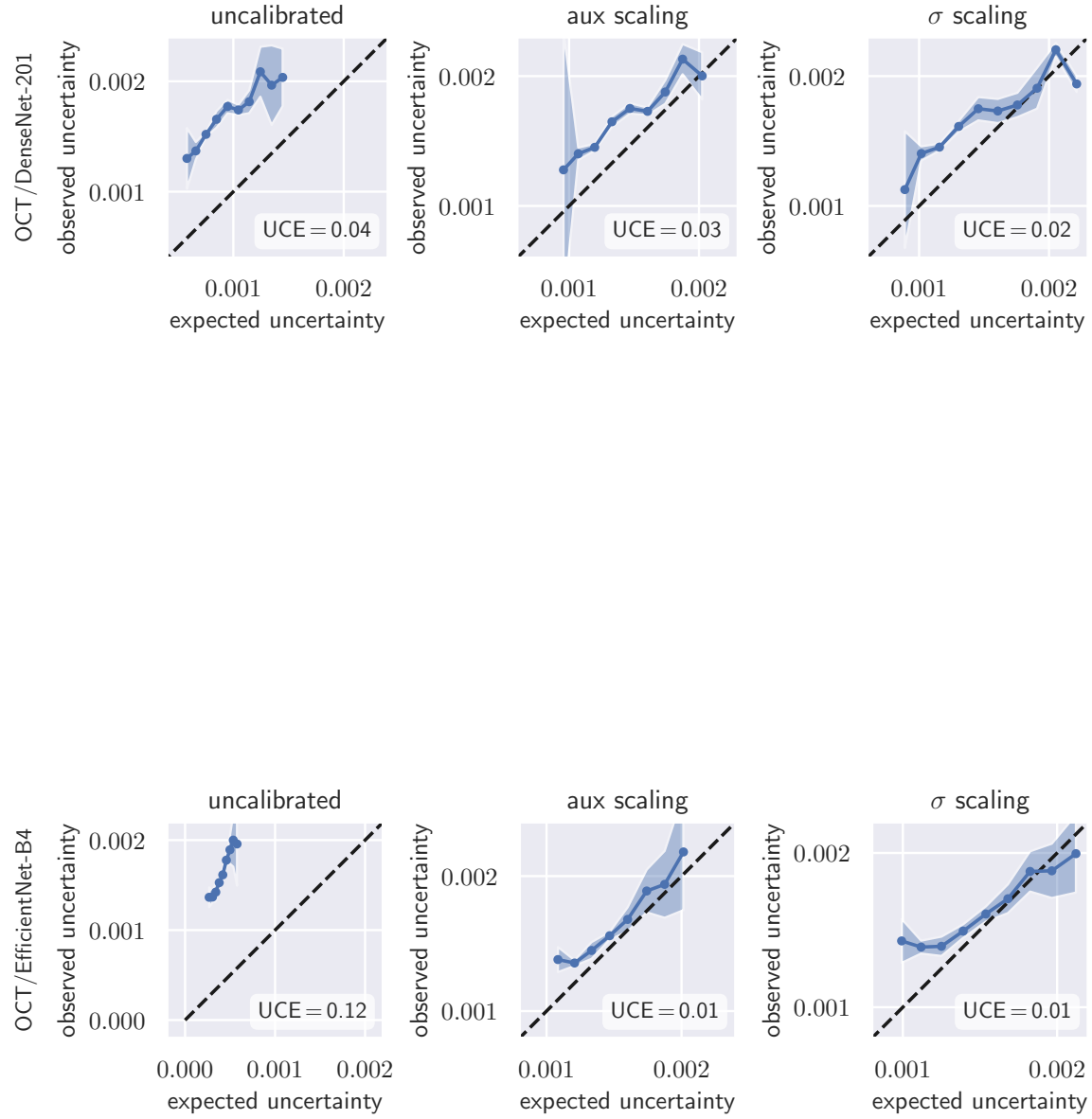
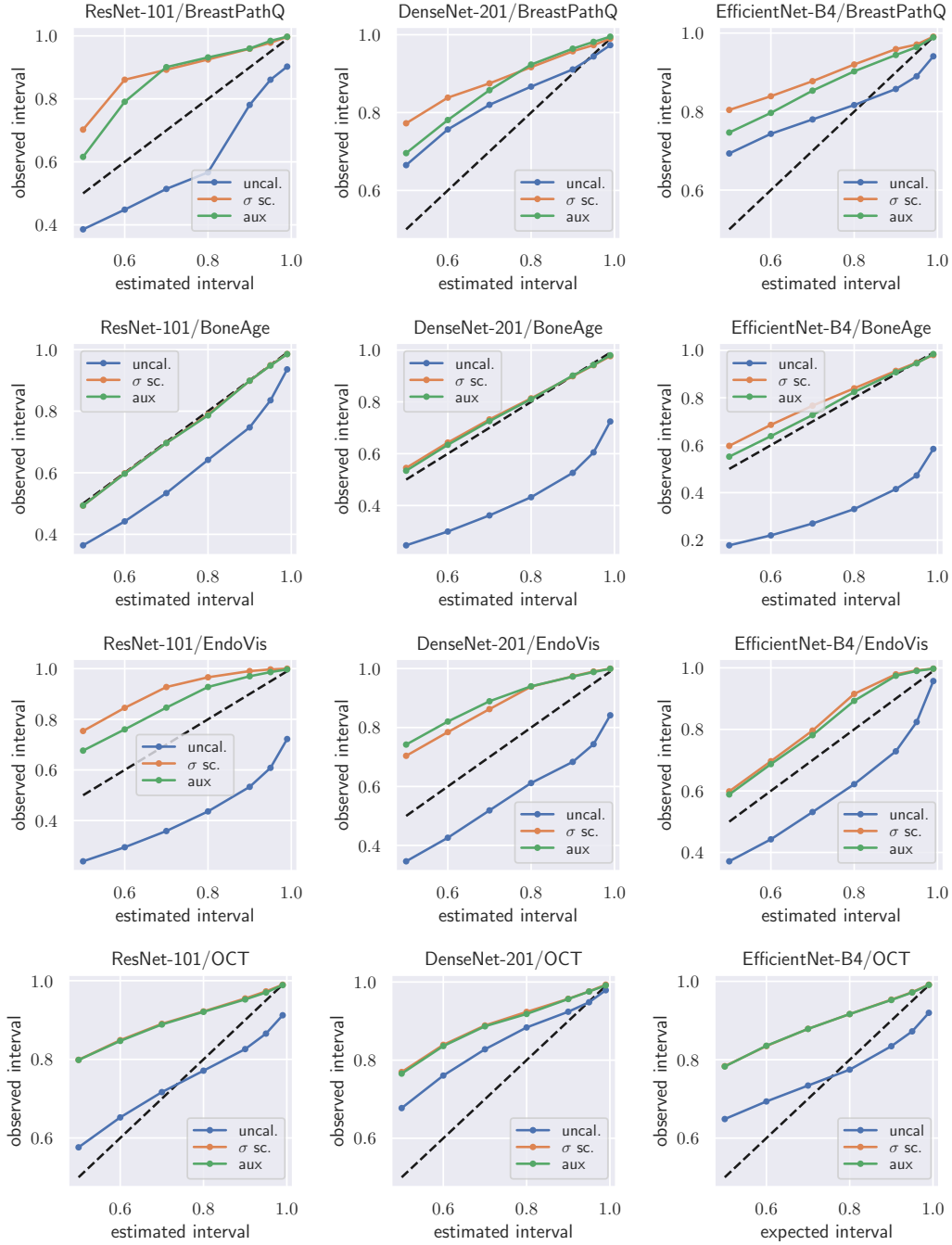## G.1 Additional Prediction Intervals

Figure 11: Observed vs. estimated posterior prediction intervals for all networks/data sets.