Uncertainty Quantification in Medical Image Synthesis

Riccardo Barbano*† Simon Arridge†‡ Bangti Jin‡
Ryutaro Tanno†§

June 18, 2021

Abstract

Machine learning approaches to medical image synthesis have shown outstanding performance, but often do not convey uncertainty information. In this chapter, we survey uncertainty quantification methods in medical image synthesis and advocate the use of uncertainty for improving clinicians' trust in machine learning solutions. First, we describe basic concepts in uncertainty quantification and discuss its potential benefits in downstream applications. We then review computational strategies that facilitate inference, and identify the main technical and clinical challenges. We provide a first comprehensive review to inform how to quantify, communicate and use uncertainty in medical synthesis applications.
Key words: uncertainty quantification; medical image synthesis; deep learning; approximate inference; Bayesian neural networks

1 Introduction

Machine learning has had a pivotal impact on medical image synthesis, which describes the task of synthesising an image of a target modality. In this chapter, we adopt a generic definition of the task in order to encompass both traditional synthesis problems of generating images from available ones of different modalities [1, 2, 3, 4, 5], and reconstruction problems in which the creation of images is performed from raw acquisition data[1]. Figure 1 illustrates this categorisation and the far-reaching impact of machine learning methods in a number of synthesis applications.

*Department of Medical Physics, UCL, Gower Street, London WC1E 6BT, UK
†Centre for Medical Image Computing, UCL, Gower Street, London WC1E 6BT, UK
‡Department of Computer Science, UCL, Gower Street, London WC1E 6BT, UK
§Healthcare Intelligence, Microsoft Research Cambridge, UK
[1]This class of problems are usually referred to as *inverse problems* in imaging. From a statistical standpoint, an inverse problem can also be interpreted as a generating process [6, 7].

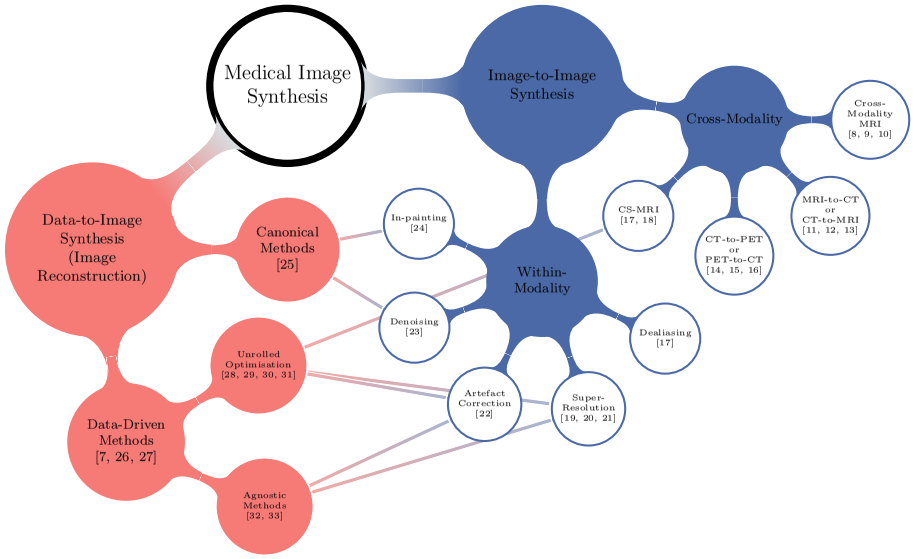
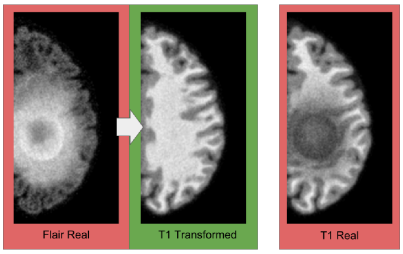
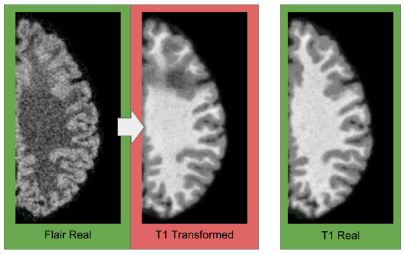


Figure 1: The diagram illustrates our categorisation of synthesis problems. We split the subject into two main branches. While operating in an image domain (i.e. image-to-image synthesis), we identify two sub-categories: (i) within-modality image synthesis (e.g. denoising, super-resolution, dealiasing, artefact correction); (ii) cross-modality image synthesis, focusing on knowledge transfer across medical image modalities. For example, one may want to learn a mapping to translate image intensities between either two Magnetic Resonance Imaging (MRI) contrasts or two image modalities: cross-modality Magnetic Resonance Imaging (MRI) (e.g. T1 weighted, T2 weighted, fluid-attenuated inversion recovery, and Magnetic Resonance Angiography), Magnetic Resonance Imaging (MRI)-to-Computed Tomography (CT) or Computed Tomography (CT)-to-Magnetic Resonance Imaging (MRI), Compressed Sensing (CS)-MRI, Computed Tomography (CT)-to-Positron Emission Tomography (PET) or Positron Emission Tomography (PET)-to-Computed Tomography (CT). While operating in a data acquisition domain (and data denotes the measurements collected from a given imaging modality) we identify a wide category of data-to-image synthesis. This is also referred to as inverse problems in medical imaging, or image reconstruction. We then identify two sub-groups: (i) canonical model-based regularisation methods (e.g. variational and iterative) and (ii) data-driven methods, including unrolled optimisation methods (physics-guided) and agnostic domain-transform methods (purely data-driven). For each sub-category we highlight a few exemplary applications as small white circles. Some applications overlap between the two groups and this is illustrated by linking them to the relevant sub-categories in the respective groups.

As machine learning applications in image synthesis progress towards clinical translation, the question of their safety at the "bedside" becomes paramount [34, 35]. In particular, deep learning methods [36], which have recently demonstrated great promise in image synthesis (see Figure 1), often produce unexpectedly erroneous results in deployment domains when they deviate from the training one. Cohen et al. [34] provide several examples of such catastrophic failures in which the deep learning synthesis model overfits to biases in the training data and, as a result, either removes an existing focal pathology (e.g. lesions, tumours, etc) or hallucinates spurious ones (see Figure 2), rendering the outputs unusable for subsequent clinical decisions. More recently, Antun et al. [35] have shown that well-established deep learning approaches to undersampled MR reconstruction are unstable under small perturbations to the input data (see Figure 3). To make matters worse, such unreliable predictions are often perceptually realistic, thus increasing the risks of letting such failures go undetected and slip into the hands of clinicians. So long as the instability of machine learning models remains a challenge in image synthesis, we will be in need of an effective means to quantify the risks of failures and to ultimately prevent failures from arising.



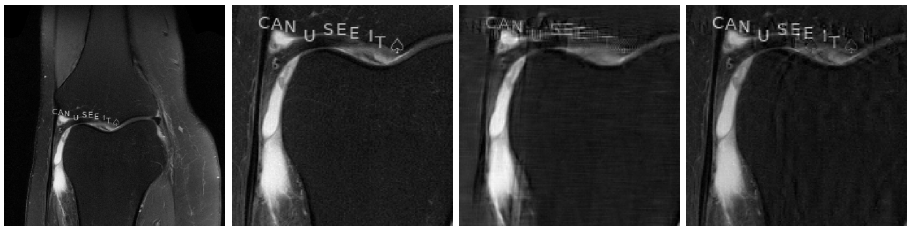
(a) A translation removing tumours



(b) A translation adding tumours

Figure 2: Examples of failures under data shifts in deep learning based FLAIR-T1 MR synthesis. Images of healthy subjects and those with tumours are shown in green and red. In (a), the model is only trained on images of healthy subjects and as a result ends up removing a tumour in the test domain. In (b), the model is trained only on images of tumour patients and tested on healthy cases, leading to the creation of a synthetic tumour, which is not present in the original image. Source: [34].

It has been argued that uncertainty quantification provides a powerful framework to address this challenge [41]. So far, the overwhelming majority of methods in synthesis (mainly, machine learning based ones but also others) deliver a single prediction, but leave users with no measure of its reliability. The ramifications and the forms of synthesis failures depend on the specifics of the downstream processing and the decision-making that consumes the synthesised images. This necessitates quantifying the risks of using synthesised images in a way that is tailored to its clinical end use. Furthermore, the users may desire to understand the sources from which the risks originate (e.g. the test case is



(a) Target (b) Target (Zoomed) (c) Deep Learning (d) Compressed Sensing

Figure 3: Examples of instabilities produced by neural networks for undersampled Magnetic Resonance Imaging (MRI) reconstruction. In (a), small structured perturbations (in the form of text and symbols) are introduced (e.g. "CAN YOU SEE IT ♠"). In (c) and (d), the reconstructions from MRI Variational Network (MRI-VN) [37] and state-of-the-art classical methods (i.e. compressed sensing [38, 39, 40]) are shown, respectively. MRI-VN is moderately unstable with respect to structural changes; such instability coincides with the inability to reconstruct details. Note that MRI-VN has not been trained with images containing any of the letters or symbols used in the perturbation. Source: [35].

under-represented in the training data vs. inherently ambiguous), so they can act accordingly to mitigate them. Uncertainty quantification allows us to formalise these practical challenges in the language of probability theory and to design potential solutions [42, 43]. While the wider machine learning community has begun to realise the importance of quantifying uncertainty information [44], this topic has yet to receive the attention it deserves in image synthesis.

The scope of this chapter is to identify current and future challenges in uncertainty quantification for medical image synthesis along with possible uses in clinical practice. Throughout the chapter, although primarily focusing on uncertainty quantification in deep learning methods, we survey "classical" approaches (i.e. approaches developed prior to the advent of deep learning), because many of the concepts we cover are generally applicable to machine learning approaches. We also discuss modelling challenges from the standpoint of machine learning developers. We discuss whether uncertainty information should be directly communicated to clinicians or used as a part of the background safety mechanism within the system. Furthermore, we query to what extent risk management should depend on the specific synthesis task of interest and its downstream usage in practice. For example, the diagnosis of different conditions and different deployment environments (e.g. A&E vs standard practice) may require synthesised images of different quality and hence different degrees of reliability.

In this chapter, we provide the first comprehensive review of uncertainty quantification in medical image synthesis. Moreover, we highlight the main research gaps and foreseeable challenges. The rest of the chapter is structured as follows. In §2 we provide background on uncertainty quantification. In §3 we discuss traditional and deep learning approaches for handling uncertainty. Lastly, in §4 we discuss the technical (and practical) challenges associated with quan-

tifying uncertainty, and the obstacles in translating uncertainty-aware methods into clinical practice.

2 Troublesome Uncertainty Landscape

Uncertainty quantification has recently begun to attract attention in the medical imaging community [45, 46, 47, 48, 49][2]. To date, however, the subject remains severely under-explored for image synthesis applications[3].

This section is structured as follows. In §2.1 we attempt to answer the question: "What is uncertainty quantification?", and we present a taxonomy of uncertainty with an emphasis on the distinction between aleatoric and epistemic uncertainty. In §2.2 we motivate why we should care about quantifying uncertainty. Lastly in §2.3 we propose a case study where we exemplify how an existing synthesis framework may benefit from uncertainty quantification.

2.1 What Is Uncertainty Quantification?

Imagine you were given a machine learning model $\mathrm{F}_\theta(\cdot)$ that takes a query instance x_q (e.g. an input magnetic resonance (MR) image) and makes a prediction $\hat{y}_q = \mathrm{F}_\theta(x_q)$ about a target image of interest y_q (e.g. a computed tomography (CT) image), where y_q and $\hat{y}_q$ denote the target output variable and its estimate from the model $\mathrm{F}_\theta(\cdot)$, respectively. The model $\mathrm{F}_\theta(\cdot)$ is parametrised by a (possibly high-dimensional) vector θ, which is optimised based on the training dataset consisting of N pairs of inputs and target outputs $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^N$. In a supervised learning setting, we assume the existence of some θ that controls the dependence between the input and output $p(\mathcal{D}|\theta)$ (i.e. the likelihood of θ). Synthesising a Computed Tomography (CT) from an MR image is a problem of predictive inference: given a set of data $\mathcal{D}$ and a query x_q, what is the associated prediction $\hat{y}_q$? In the framework of probabilistic machine learning, inference involves several learning and approximation steps, and all the errors and uncertainties incurred at these steps contribute to the uncertainty of the output $\hat{y}_q$. Below, we present a taxonomy of different uncertainty types and explain their differences and interrelations (see Table 1).

Predictive uncertainty is a measure describing the degree of ambiguity (or confidence) in the model's output $\hat{y}_q$ for a given input x_q. For example, we may report the 95% confidence interval for each pixel (i.e. capturing two standard deviations on either side of the mean estimate, under the Gaussian

[2]See Abdar et al. [44] for a comprehensive review on uncertainty-aware methods in deep learning; medical image classification, segmentation and registration are also thoroughly discussed.

[3]Although under-explored for image synthesis applications, uncertainty quantification is an important, ongoing research topic within the machine learning community, and for instance, just recently complementary yet alternative formal definitions (to the ones we provide in this chapter) on model bias, model variance, and aleatoric and epistemic uncertainty have been proposed [50].

Table 1: Uncertainty types and their distributional forms. Model M denotes one element from model class $\mathcal{M}$, e.g. a neural network $\mathrm{F}_\theta(\cdot)$ with the associated parameter vector θ.

Uncertainty Type	Distributional Form	Ambiguity in
Predictive	$p(\hat{y}\|x)$	the model's output
Aleatoric	$p(y\|x)$	the data formation process
Epistemic - Structural	$p(\mathrm{M}\|\mathcal{D})$	the model specification
Epistemic - Parametric	$p(\theta\|\mathcal{D})$	the estimation of the model parameters

assumption) along with the synthetic image as a measure of predictive uncertainty. The confidence interval can then be used to assess the variability of the prediction (e.g. the smaller the interval, the more certain the model is about the prediction). Predictive uncertainty is represented in, what is known as, the (posterior) predictive distribution $p(\hat{y}_q|x_q)$.

One is often interested not only in quantifying predictive uncertainty, but also in understanding its sources [51, 52], which are useful in identifying the factors from which predictive uncertainty arises. In medical image synthesis, Tanno et al. [53] have shown how disentangling the constituents of uncertainty yields a form of interpretation of predictive uncertainty. The sources of predictive uncertainty are typically further categorised into *aleatoric* and *epistemic* uncertainty [54, 55, 56, 57, 58].

Aleatoric uncertainty — from the Latin word *alea* meaning a die — refers to the uncertainty inherent to a problem or an experimental setup that cannot be reduced by additional physical or experimental knowledge [59]. It is also referred to as data, intrinsic or irreducible uncertainty in collected measurements caused by the presence of stochasticity (e.g. measurement noise [60], data transmission and storage errors). For instance, when synthesising CTs from MR images, aleatoric uncertainty stems from the fact that there are multiple plausible CT solutions for a single MR image. Uncertainty of aleatoric nature is summarised by the underlying conditional distribution $p(y_q|x_q)$ of the task, which describes the inherent stochasticity in the system output y_q for the given input x_q. Such uncertainty is irreducible by collecting more data under experimental settings. If we wish to reduce aleatoric uncertainty (e.g. the noise in the acquired data), we might have to switch to a different acquisition protocol.

Epistemic uncertainty — from Ancient Greek "$\epsilon\pi\iota\sigma\tau\eta\mu\eta$" meaning knowledge — refers to the uncertainty arising from a lack of knowledge or statistical evidence (i.e. the "epistemic" state of the decision maker). It is often further decomposed into two sources: *structural* and *parametric* uncertainty.

Structural uncertainty (or model inadequacy) refers to the uncertainty about whether the model is structurally correct. It is also referred to as model specification uncertainty or architecture uncertainty [61]. In fact, we might even be uncertain about whether we have chosen the correct model class in the first

place. Perhaps, the current model class does not explain the data well, and if it is inadequate, we may need to construct a different one. It is expressed as the plausibility of the true target process to lie in the specified model class $\mathcal{M}$. It is thus described by a distribution $p(\theta \in \mathrm{M}|\mathcal{D})$ which quantifies how probable it is that model M (e.g. a neural network (NN) $\mathrm{F}_\theta(\cdot)$ with the parameter vector θ) is within the model class $\mathcal{M}$, given the data $\mathcal{D}$. Model uncertainty is strictly related to multi-model inference [62], which subsumes Bayesian model comparison, selection and averaging, as there may exist a multitude of model classes that explain the data equally well. Is linear regression appropriate? Or a neural network? If the latter, how many layers should it have? In medical image synthesis, we often assume that the hypothesis space Θ is correctly specified and neglect the risk of model misspecification.

Parametric uncertainty denotes the uncertainty related to the estimation of the model parameters under a given model specification, assuming that the form of the model faithfully captures reality. Consider a scenario in which we choose a complex model (with ≈ 60 million parameters) but we lack a sufficient amount of training data (as is often the case in medical image synthesis) to train our model on. In this case, we will likely struggle to constrain the model's parameters. Out of all the "functions" our model can represent, which one should we choose? Parametric uncertainty is described by the posterior distribution $p(\theta|\mathcal{D})$ over the unknown parameters θ of the specified model $\mathrm{F}_\theta(\cdot)$, given the data $\mathcal{D}$. The more "peaked" $p(\theta|\mathcal{D})$ is (i.e. the more concentrated the probability mass is in a small region in Θ), the less uncertain the decision maker should be. In other words, high parametric uncertainty arises when the predictions obtained from several "plausible" parameter settings disagree the most.

Many technical and practical problems with uncertainty quantification boil down to estimating these distributions in various settings. For complex models such as NNs, these distributions are mostly intractable, necessitating the development of efficient and effective approximations. In medical imaging synthesis, the "ground truth" for these distributions of interest, $p(\hat{y}|x_q)$, $p(y|x_q)$, $p(\theta|\mathcal{D})$, and $p(\mathrm{M}|\mathcal{D})$ are often not explicitly available, rendering the exact evaluation of uncertainty estimation very challenging [63]. We shall describe efficient strategies for tractable approximations in §3.

2.2 Why Should We Care?

Uncertainty quantification offers a principled and consistent framework that provides reliability measures of the model's output, which potentially can shed valuable insight for downstream applications. In this regard, we argue that uncertainty quantification could assist the translation of medical image synthesis technologies into clinical practice while improving clinicians' trust [64]. Below we present four use cases of estimated uncertainty information in a variety of settings: quality check, propagating uncertainty, shedding insight and improving predictive performance. We also present the safety implications of deploying machine learning based image synthesis applications in clinical practice.

Quality Check

Taking contrast enhancement of CTs as an exemplary synthesis application, one may be interested in whether the model generalises in new environments. One may want to asses if the model can reliably enhance the CTs of all relevant sub-populations that are not well-represented in the training data. Or, one may want to know how the model would behave if the acquisition parameters of the CT scanner or even its type were to change in the deployment site due to some operational reasons. How would the model perform if the CTs of patients with rare conditions or diseases were to be taken? Ideally, we would collect enough validation data in all these possible scenarios and assess the model's performance. Such an approach, however, is impractical. To make matters worse, several works have shown that deep learning models often overestimate their confidence in the synthesis process. First, Cardoso et al. [4] warn about the "risks" of the model being overconfident, and possibly propagating large errors to downstream analysis. Then, Cohen et al. [34] and Antun et al. [35] warn about the dangers of machine learning models hallucinating image features, and advocate the need for a quality check for image-to-image translation and MR image reconstruction.

To address these questions, we can look at recent works. Tanno et al. [53] have suggested that predictive uncertainty, if quantified correctly, provides a surrogate performance metric that could reliably inform the clinicians when not to trust the model's predictions. They propose a Bayesian image quality transfer via convolutional neural networks (CNNs) [65] and demonstrate the usefulness of uncertainty modelling by measuring the deviation from the ground truth on standard metrics. The standard deviation map highly correlates with reconstruction errors, which shows their potential as a surrogate measure of accuracy. More recently, Tanno et al. [66] show that predictive uncertainty can be used to define a binary classifier, discriminating "risky" predictions from the "safe" ones. In a different synthesis task, Reinhold et al. [67] propose a Bayesian deep learning method that learns how to translate a CT into an MR image and to quantify uncertainty, which is then used as a proxy for anomaly detection. On the basis that high pixel-wise uncertainty occurs in pathological regions of the synthetic CT, Reinhold et al. [68] use uncertainty quantification for unsupervised anomaly segmentation. Klaser et al. [69] propose a novel multi-resolution cascade 3D network for end-to-end full-body MR to CT synthesis yet include uncertainty quantification as a measure of safety. Lastly, Nair et al. [46, 70] investigate several uncertainty metrics for quality control in lesion segmentation of multiple sclerosis.

Propagating Uncertainty

Clinical researchers may use predictive uncertainty in downstream analysis, or include it in the pipeline of medical image analysis, which generally comprises a sequence of inference tasks (e.g. synthesis, registration and segmentation). The uncertainty quantified at the image level is passed to subsequent tasks

in the form of an uncertainty map (e.g. pixel-wise variance). Recent works have explored this prospective use. Tanno et al. [53] propagate uncertainty into downstream quantities in the context of diffusion MRI super-resolution, by computing the expectation and variance of mean diffusivity and fractional anisotropy with respect to the predictive distribution. Mehta et al. [71] show how the performance of a downstream task in a medical image analysis pipeline can be improved if uncertainty estimates are propagated: the output of each module (including the associated uncertainty) is used as an input to the subsequent one across cascaded inferential tasks. The paper studies several medical image pipelines, each of which cascades two different inferential tasks (e.g. two-stage Magnetic Resonance Imaging (MRI) synthesis and brain tumour segmentation). Experimental results indicate that propagating the synthesised image along with its associated uncertainty map to the downstream tumour segmentation network improves the downstream performance in comparison to only propagating the synthesised image.

Shedding Insight

In a scenario where the synthesis error is consistently high on certain image structures, decomposing predictive uncertainty into aleatoric and epistemic uncertainty provides high-level "explanations" for a model's behaviour. For instance, such a decomposition allows quantifying how much uncertainty arises from (i) the inherent difficulty to reconstruct image structures (i.e. uncertainty of aleatoric nature); (ii) the unfamiliarity of such image structures due to their limited representation in the training data (i.e. uncertainty of epistemic nature). If the epistemic uncertainty is high but the aleatoric one is low, this indicates that collecting more training data would be beneficial. On the contrary, if the epistemic uncertainty is low and the aleatoric one is high, then we need to regard such errors as inevitability, and abstain from predictions to ensure safety or account for errors appropriately in subsequent analysis. Data-driven approaches for uncertainty quantification also present an additional technical challenge: the selection and collection of the training data and the evaluation of its completeness and accuracy. Disentangling the constituents of predictive uncertainty may suggest how to collect the training data, and the extent to which it is informative and exhaustive. Tanno et al. [66] show that the decomposition of the effects of aleatoric and epistemic uncertainty in the predictive uncertainty provides additional explanations of the performance of the considered methods.

Improving Predictive Performance

Bayesian approaches to machine learning models offer a number of theoretical as well as practical advantages. They provide a potential solution to over-fitting, and a principled and automatic way of selecting hyper-parameters [63, 72, 73]. Many techniques of regularisation arise in a natural way in the Bayesian framework as the maximum a posteriori (MAP) estimator of certain posterior probability density functions. The need for regularisation is compelling in the con-

text of deep learning based techniques, where nearly all models are severely over-parameterised, due to a lack of abundant high-quality training samples. Bayesian approaches also deliver quantifiable estimates of uncertainty of the model parameters and predictions as well as quantitative comparisons between predictions obtained by alternative models (e.g. different network architectures) within the framework of model selection (e.g. using Bayes factor [74]). Furthermore, these approaches enable developing "optimal" estimators with respect to suitable Bayesian risks within the Bayesian decision theory framework [75].

In order to fully realise these advantages, there remain computational challenges, which we shall discuss in detail in Section 3.

2.3 Uncertainty Quantification in Action

Lastly, we would like to end this section by illustrating how uncertainty could be used in image synthesis applications. In fact, there are many scenarios in which uncertainty quantification could be useful to clinicians. Here, we illustrate how positron emission tomography (Positron Emission Tomography (PET))/MR image reconstruction may benefit from uncertainty-aware attenuation correction in PET. Clinical researchers have improved PET/MR reconstruction by generating a "pseudo-CT" and deriving the attenuation coefficients [3], which, in turn, play a substantial role in PET reconstruction. The synthetic information is implicitly used within the reconstruction pipeline to inform the attenuation coefficients, and it is also customary for nuclear medicine physicians to visualise the pseudo-CT for PET/MRI (CT in case of PET/CT) mainly to check the movement artefact. In theory, one should check the plausibility of pseudo-CTs, which is however rarely done in practice. Note that by detecting obvious artefacts (e.g. air in the middle of the brain because of a segmentation problem), one may find the wrong bone density. What happens if the approach is unable to correctly synthesise a CT from the MR image? This might be the case for patients that have evident bone defects (e.g. low or high bone density). For such an "outlier" patient, a notion of uncertainty over the pseudo-CT could be useful as it would provide a background defensive mechanism that informs the clinician not to use the pseudo-CT and attenuation maps as "too risky" to trust for PET reconstruction. We may want our automated system to abstain from using the pseudo-CT and request the assistance of a clinician when the uncertainty is above a certain threshold.

3 Tools for Modelling Uncertainty

Now we turn to practical computational strategies for handling uncertainties within the Bayesian framework [76, 77]. The main idea is as follows: all the quantities, which appear in synthesis tasks are modelled probabilistically as random variables with corresponding probability distributions (e.g. density for continuous random variables). Within the Bayesian framework, there are two fundamental building blocks: the likelihood (of the training data $\mathcal{D}$) and the

prior distribution. The training data set $\mathcal{D}$ consists either of a set of available measurements y in data-to-image synthesis or a set of N pairs (x_i, y_i) in the context of supervised learning. The prior distribution $p(\theta)$ of the parameter θ specifies the prior knowledge we have before collecting the measurements. In the context of standard data-to-image synthesis, θ is the target image and $p(\theta)$ encodes the *a priori* knowledge we have about the sought-for image, whereas in supervised learning, we seek to learn the posterior distribution $p(\theta|\mathcal{D})$ over the parameters θ of the model $\mathrm{F}_\theta(\cdot)$. Learning consists of updating the prior distribution $p(\theta)$ to the posterior distribution $p(\theta|\mathcal{D})$ defined as:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)\mathrm{d}\theta}, \tag{1}$$

where the likelihood function of parameters θ, $p(\mathcal{D}|\theta)$, is the probability of the given training data set $\mathcal{D}$ given θ. The posterior distribution $p(\theta|\mathcal{D})$ over θ is inferred by *deductively* updating the prior knowledge $p(\theta)$ we had, given the data $\mathcal{D}$ we observed [78, 79]. Note that, in "machine learning parlance", we usually denote the input by x and the target output by y, whereas in the inverse problem community, y denotes the observations (i.e. measurements that have undergone through a corruption process) and x is the image to be reconstructed (and θ, for instance, is the parameter vector of the neural network). Here we will follow the machine learning notation.

To represent uncertainty about a prediction $\hat{y}_q$ all possible configurations of θ are considered, with each prediction being weighed by its posterior probability $p(\theta|\mathcal{D})$. Thus, we compute the posterior predictive distribution $p(\hat{y}_q|x_q)$:

$$\underbrace{p(\hat{y}_q|x_q)}_{\substack{\text{Predictive}\\ \text{Uncertainty}}} = \int \underbrace{p(\hat{y}_q|x_q,\theta)}_{\substack{\text{Aleatoric}\\ \text{Uncertainty}}} \underbrace{p(\theta|\mathcal{D})}_{\substack{\text{Epistemic}\\ \text{Uncertainty}}} \mathrm{d}\theta, \tag{2}$$

which captures both aleatoric and epistemic uncertainty. The final prediction is obtained by Bayesian *model averaging*; or if stated differently, is made through Bayesian *marginalisation* as the predictive distribution of interest no longer conditions on θ. Intuitively, we can think of Eq. (2) as a weighted average (i.e. the outcome of a reconstructed image) of many different hypotheses by their plausibility given data — we would like to use every possible setting of θ — rather than a single one. In the supervised learning setting, the challenges in computing the posterior predictive distribution $p(\hat{y}_q|x_q)$ are two-fold: (i) estimating the posterior distribution $p(\theta|\mathcal{D})$; (ii) integrating out θ. Since Bayesian model averaging is often too hard, we either tend to approximate the integral with a simple Monte Carlo (MC) approximation:

$$p(\hat{y}_q|x_q) \approx \frac{1}{T}\sum_{t=1}^{T} p(\hat{y}_q|x_q,\hat{\theta}^t), \quad \text{with } \hat{\theta}^t \sim p(\theta|\mathcal{D}),$$

or, we adopt only the single prediction with the highest posterior distribution:

$$\theta_{\text{map}} = \operatorname{argmax} p(\theta|\mathcal{D}). \tag{3}$$

This estimate is commonly known as the MAP estimate and is computationally more tractable. Even though MAP involves the posterior distribution $p(\theta|\mathcal{D})$ and looks like an application of the Bayes' rule, it is not properly Bayesian. In fact, it would put everything on one single hypothesis, that is, on a single setting of the parameters $\mathrm{F}_{\theta_{\mathrm{map}}}(\cdot)$. Accordingly, Eq. (2) would be computed by using an approximate posterior distribution $p(\theta|\mathcal{D}) \approx \delta(\theta = \theta_{\mathrm{map}})$, where δ is a Dirac delta distribution with all its mass at θ_{map}, with the likelihood being $p(\hat{y}_q|x_q, \theta_{\mathrm{map}})$. The difference between these two approaches relies on the posterior distribution $p(\theta|\mathcal{D})$, but most importantly on how "sharp" it is. In fact, there would be almost no difference if the posterior distribution happened to be sharply peaked, and the likelihood $p(\hat{y}_q|x_q, \theta)$ did not vary much in the region where the posterior distribution places its mass. A Dirac delta may then be a reasonable approximation of the posterior distribution in Eq. (2)[4]. If this is not the case, averaging the predictions of many high performing models $\hat{\theta}^t$ (e.g. neural networks) that "disagree" for some input cases can lead to a significant improvement in accuracy [82, 81].

3.1 Approximation Techniques

Although the posterior distribution $p(\theta|\mathcal{D})$ gives a complete probabilistic solution to the synthesis task — it combines both the prior knowledge with the given data — a closed form expression for $p(\theta|\mathcal{D})$ is often unavailable in medical image synthesis. There are several forms of intractable posterior distributions: (i) the normalising constant is intractable (i.e. "analytically" intractable); (ii) the posterior distribution is intractable due to an intractable likelihood (e.g. the data generating process being too complex due to poorly understood physics). Generally, summary statistics (e.g. mean and variance or correlation) are sought. However, these quantities require computing high-dimensional integrals, which are computationally infeasible for most synthesis tasks. Thus, it is imperative to employ numerical procedures to effectively explore the posterior distribution $p(\theta|\mathcal{D})$. These can roughly be divided into two groups: MC type methods and approximate inference techniques. MC type methods include Markov chain Monte Carlo (MCMC), which constructs a Markov chain whose stationary distribution is the posterior distribution, and which uses ergodic averages to approximate the statistics of interest, and sequential MC, which constructs a finite sequence of importance samplers targeting a sequence of distributions with the last being the posterior distribution.

Approximate inference methods include the Laplace approximation using a local Gaussian approximation constructed at the MAP, Variational Inference (VI) framing the approximation of the posterior distribution as optimising a lower bound on the evidence with respect to a tractable family of simple distributions (commonly referred to as a variational distribution), and expectation propagation, which iteratively leverages the factorisation structure of the target

[4]However, this is hardly the case for neural network, which are highly under-specified by the available data [80, 81].

distribution.

Before we proceed further, it is useful to recall that the end goal is to accurately approximate the posterior predictive distribution in Eq. (2). To do so, it is important to have an accurate approximation of the posterior distribution in the regions that would contribute most to the Bayesian model averaging integral in Eq. (2). Let's imagine one samples two different settings of parameters of the network $\mathrm{F}_\theta(\cdot)$, namely $\hat{\theta}^1$ and $\hat{\theta}^2$, but both give rise to similar functions $\mathrm{F}_{\hat{\theta}^1}(\cdot)$ and $\mathrm{F}_{\hat{\theta}^2}(\cdot)$. In this case, the second setting of parameters $\hat{\theta}^2$ would not contribute much to estimating the integral in Eq. (2), and we should seek functional diversity for a good approximation of Eq. (2) [83].

Monte Carlo Methods

In MC type methods, one generates samples from $p(\theta|\mathcal{D})$, which are then used to produce representative reconstructions or to compute summary statistics. Directly generating samples is generally very challenging. MCMC [84] methods (e.g. the Metropolis-Hastings algorithm or the Gibbs algorithm) generate a Markov chain whose stationary distribution is the target distribution, and is asymptotically exact. These methods can approximate the target distribution arbitrarily well, provided that one can run the chain for sufficiently long, and thus have been established as the gold standard for exploring the posterior state space. In practice, these methods are often easy to implement, but their efficiency relies heavily on various algorithmic parameters (e.g. proposal distribution and step-size). To make matters even worse, the scalability with parameter dimensionality is often not very favourable and the convergence diagnosis remains largely an art rather than a science.

Consequently, despite their impressive progress in recent years (e.g. Hamiltonian Monte Carlo [85]), the use of MC methods in the context of medical image synthesis (including image reconstruction) remains fairly limited. However there are some exceptions. Pedemonte et al. [86] use a recent Riemann manifold MCMC sampling scheme [85] to sample the posterior distribution of emission rates given the photon counts for PET. The method obtains uncertainty information from all the processes involved in the reconstruction algorithm (i.e. the observed data, the measurement noise and the background signal, the reconstruction process itself, and also possibly the hyperparameters). Moreover, the tightening of the posterior distribution is also used as a reliability indicator for estimating the required patient scan time. Weir et al. [87] propose an approach for SPECT that samples the joint posterior distribution of the image and hyperparameters using a Gibbs prior and the Metropolis-Hastings sampler on simulated and phantom data. Similarly, Barat et al. [88] propose a Gibbs sampler for PET with a nonparametric Dirichlet process mixture prior. However, even for medium-size medical image reconstruction, exploring the posterior distribution with MCMC type methods can incur a prohibitively high computational expense, and thus is not practically feasible. As a rule of thumb, the higher the dimensionality, the more complex the posterior distribution is, and the slower the sampling procedure converges. For PET, Filipovic et al. [89]

develop a Gibbs type sampler formed from a distance-driven Chinese restaurant process (for clustering). Nonetheless, the procedure remains expensive: "The computation time was 4 days for RCP-GS (30 runs × 250 sampler iterations), compared to 1h20 for MR-MAP and 50min for OSEM (8 iterations × 27 subsets)" [89].

Approximate Inference Schemes

Deterministic approximate inference techniques encompass a large variety of methods such as the Laplace approximation [90], VI [91, 92, 93] (using mean-field approximation [94], or the variational Gaussian approximation [95, 96] and more recently stochastic VI [97]), and expectation propagation [98].

The Laplace approximation is a classical approach to approximate the posterior distribution. It constructs a Gaussian distribution based on the second-order Taylor expansion of the log-posterior $\log p(\theta|\mathcal{D})$ around the MAP estimate θ_{map}:

$$p(\theta|\mathcal{D}) \propto \exp\left\{-\frac{1}{2}\left(\theta-\theta_{\text{map}}\right)^{\top} \mathrm{H}(\theta_{\text{map}})\left(\theta-\theta_{\text{map}}\right)\right\}, \tag{4}$$

where $\mathrm{H}(\theta_{\text{map}}) = -\nabla^2_\theta \log p(\theta|\mathcal{D})|_{\theta=\theta_{\text{map}}}$ denotes the Hessian of the (negative log) posterior distribution estimated at the MAP estimate θ_{map}. This approach requires good differentiability of the negative log posterior distribution, and it is thus not directly suitable for non-smooth priors (e.g. sparsity or total variation) which commonly appear in image reconstruction[5]; but most importantly, computing the full Hessian H_θ is computationally demanding and memory-wise infeasible, unless further fast approximations (e.g. diagonal + local rank, Kronecker or sparse (inverse) covariance) are employed. The low-rank assumption is reasonable for severely ill-posed imaging problems. It is also worth noting that often more accurate approximations can be obtained using the integrated nested Laplace approximation [100]. Despite its simplicity, it has barely been employed in medical image restoration or synthesis.

Most VI techniques were developed within the machine learning community, where the aforementioned computational challenge is widely acknowledged. VI is often based on approximately minimising the Kullback-Leibler (KL) divergence[6] [101] between the target distribution and the approximate surrogate one. The divergence KL from q to p is defined by

$$\mathrm{KL}(q\|p) = \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x. \tag{5}$$

VI then searches for an approximating distribution $q^*_\psi(\theta)$ parametrised by ψ

[5] Various smoothing (e.g. Huber) can be used for non-smooth priors, but it can also significantly hinder the approximation; see [99] an illustration with the anisotropic total variation prior.

[6] Note that the divergence is non-symmetric and does not satisfy the triangle inequality, but it vanishes if and only if p equals to q almost everywhere.

within an admissible family $\mathcal{Q}$ by minimising the KL divergence:

$$q_\psi^*(\theta) := \underset{q_\psi \in \mathcal{Q}}{\operatorname{argmin}} \operatorname{KL}\left(q_\psi(\theta) \| p(\theta|\mathcal{D})\right). \tag{6}$$

Introducing a prior distribution $p(\theta)$ and applying the Bayes rule allows us to rewrite the optimisation of Eq. (6) as the maximisation of the Evidence Lower BOund (ELBO) with respect to the variational parameters defining $q_\psi(\theta)$,

$$\mathcal{L}_{\mathrm{VI}} := \int q_\psi(\theta) \log p(\mathcal{D}|\theta) \mathrm{d}\theta - \int q_\psi(\theta) \log \frac{q_\psi(\theta)}{p(\theta)} \mathrm{d}\theta \leq \log p(\mathcal{D}) \tag{7}$$

The maximising functional $\mathcal{L}_{\mathrm{VI}}$ is a lower bound to the log evidence (i.e. the normalising constant or marginal log-likelihood) $p(\mathcal{D})$. Note that the ELBO plus $\operatorname{KL}\left(q_\psi(\theta) \| p(\theta|\mathcal{D})\right)$ equals the marginal log-likelihood $p(\mathcal{D})$, which is constant with respect to the variational parameters ψ.

The computational tractability of VI is achieved by imposing suitable assumptions on the approximating family $\mathcal{Q}$, for instance, a fully-factorised (a.k.a. mean-field) Gaussian $\mathcal{Q}_{\mathrm{FFG}}$ defined by

$$\mathcal{Q}_{\mathrm{FFG}} = \Big\{ q(\theta) = \prod_i \mathcal{N}(\theta_i; \mu_i, \sigma_i^2) \Big\},$$

where $\mathcal{N}(\theta_i; \mu_i, \sigma_i^2)$ denotes a Gaussian distribution for the component θ_i with mean μ_i and variance σ_i^2. The parameters μ_i and σ_i^2 are variational parameters that have to be optimised. Then maximisation is often carried out by coordinate ascent type schemes, or stochastic gradient type algorithms [97]. The latter requires an MC estimate of the gradient, which often has to be done carefully in order to ensure low bias and low variance. It is worth noting that in a different vein, suitable averaging of the stochastic gradient iterates can also be interpreted as approximate inference [102, 103], though the covariance estimate may differ in shape.

In contrast, expectation propagation [98] minimises the KL divergence defined as $\operatorname{KL}(p\|q)$, which mathematically amounts to moment matching, and its practicality relies on a factorised form of the posterior distribution and a possible reduction to low-dimensional (often still delicate) numerical integration. The stability of the implementation relies heavily on the accuracy of the quadrature rules, and an inaccurate quadrature can cause the nonconvergence of the iteration. In this regard, Zhang et al. [99] develop an approximate Bayesian inference technique based on expectation propagation for PET reconstruction (with the anisotropic total variation prior), where the delicate issue of numerical integration is studied in depth and the approach is showcased on medium-size simulated phantom data.

Besides these established approximate inference techniques, there are several others. One notable recent example is Stein Variational Gradient Descent [104], which also performs moment matching but it does so implicitly [105]. Compared with MCMC type methods, deterministic approximations are often

computationally more efficient, but may be limited in accuracy (and often with little theoretical understanding [106]), yet they remain expensive for truly large-scale problems arising in medical imaging, especially in the presence of strong correlation between different entries.

More recently, attention has also been paid to blending start-of-the-art optimisation algorithms with uncertainty quantification. For example, Repetti et al. [107] propose a method to analyse the confidence in specific structures in MAP estimates using Bayesian hypothesis testing. The method holds potential for large-scale problems, but remains to be evaluated clinically. In sum, approaches, which aim to quantify uncertainties, are mathematically principled, but there remain computational challenges; various approximations have been developed to address these challenges but a complete mathematical theory of the mathematical-statistical properties of these methods is yet to emerge and their potential in medical image analysis is to be evaluated.

3.2 Probabilistic Deep Learning

With the advent of deep learning, uncertainty quantification has resurfaced as an important framework. In medical image synthesis, Bayesian deep learning can provide the information about uncertainty associated with each prediction[7]. Below, we review the basics of Bayesian neural networks (BNNs) — a sub-field of Bayesian deep learning — which holds great potential for the image synthesis community. We also discuss how to disentangle predictive uncertainty into aleatoric and parametric uncertainty, and briefly mention several alternative approximations. Nonetheless, the aforementioned computational challenges persist due to the high-dimensionality of the parameter vector and high degree of nonlinearity within deep neural networks.

Bayesian Neural Networks

BNNs place a probability distribution on the parameters θ (which are now treated as random variables) to encode the uncertainties associated with the prediction [109, 110, 111]. We consider the posterior distribution over all possible settings of the model parameters given the observed data. Such probability density encapsulates parametric uncertainty, and its spread of mass signifies the ambiguity in selecting appropriate parameters. In recent years, there have been significant efforts to characterise and approximate the posterior distribution $p(\theta|\mathcal{D})$ [112, 113, 114], which, in practice, is intractable due to the difficult-to-compute normalising constant. It is worth noting that many approximate inference algorithms share the same approximating family $\mathcal{Q}$. For instance, VI, the diagonal Laplace approximation [115], probabilistic backpropagation [113], stochastic expectation propagation [116], black-box alpha divergence minimisation [117], Rényi divergence VI [118], natural gradient VI [119], and functional variational BNNs [120] all use a fully-factorised Gaussian family $\mathcal{Q}_{\text{FFG}}$, which itself is largely motivated by computational considerations.

[7] See Wilson [108] for a note motivating Bayesian deep learning.

We only review the two most popular schemes used in image synthesis, that is, VI and Monte Carlo dropout (MCDO), and omit to review MCMC approaches to BNNs (e.g. stochastic gradient MCMC) [121, 122, 123, 124], which remain computationally inefficient due to the evaluation of an ensemble model for the computation of the posterior distribution. We also do not review methods that construct Gaussian approximations to the posterior distribution from a few iterates along the optimisation trajectory obtained by stochastic gradient descent methods of a deterministic neural network [103]. To the best of our knowledge, these methods have yet to be applied to medical image synthesis.

VI recasts intractable inference as an optimisation problem: we replace marginalisation with the optimisation of Eq. (7), which is (unbiasedly) estimated by randomly selecting a mini-batch set $\mathcal{B}$ of M data-pairs and using $T \geq 1$ MC samples (with $\hat{\theta}^t \sim q_\psi(\theta)$) [112][8],

$$\hat{\mathcal{L}}_{\text{VI}} = \frac{N}{M} \sum_{i \in \mathcal{B}} \frac{1}{T} \sum_{t=1}^{T} \log p\left(x_i^t | y_i^t, \hat{\theta}^t\right) - \text{KL}\left(q_\psi(\theta) \| p(\theta)\right). \tag{8}$$

Currently, the most efficient technique to compute the gradients $\nabla_\psi \hat{\mathcal{L}}_{\text{VI}}$ is the so-called reparametrisation trick [125], which employs a deterministic dependence of the ELBO with respect to ψ to back-propagate. To this end, we rewrite $q_\psi(\theta)$ using a differentiable transformation $\hat{\theta}^t = g(\psi, \hat{\epsilon}^t)$ with $\hat{\epsilon}^t \sim p(\epsilon)$ and $p(\epsilon)$ being an underlying, parameter-free distribution (e.g. the standard Gaussian distribution). We can then use MC integration over $p(\epsilon)$ to evaluate the expectations, yet the value depends on θ and we can hence propagate gradients through $g(\cdot)$. The reparametrisation can be carried out either explicitly [126, 127] or implicitly [128]. Once we obtain $q_\psi^*(\theta)$ by maximising Eq. (8), we perform inference on a new query by approximating the predictive distribution in Eq. (2) as:

$$p(\hat{y}_q | x_q, \mathcal{D}) \approx \int p(\hat{y}_q | x_q, \theta) q_\psi^*(\theta) \mathrm{d}\theta := q_\psi^*(\hat{y}_q | x_q). \tag{9}$$

In practice, we approximate the optimal variational distribution $q_\psi^*(y_q | x_q)$ with MC integration:

$$\hat{q}_\psi^*(\hat{y}_q | x_q) := \frac{1}{T} \sum_{t=1}^{T} p(\hat{y}_q | x_q, \hat{\theta}^t), \quad \text{with } \hat{\theta}^t \sim q_\psi^*(\theta). \tag{10}$$

Barbano et al. [129, 130] propose a scalable and efficient framework rooted in VI formalism to jointly quantify aleatoric and epistemic uncertainties in unrolled optimisation. The framework is showcased on CT reconstruction with both sparse view and limited angle data, and the estimated uncertainty is observed to capture the variability in the reconstructions, caused by the restricted measurement model, and by missing information, due to limited angle geometry.

[8] Note that we assume that the KL term can be computed deterministically as a closed form solution might exist; otherwise it can be sampled similarly.

Gal & Ghahramani [114] propose a MCDO method, which approximates $p(\theta|\mathcal{D})$ with a multiplicative Bernoulli distribution. It defines an approximate posterior distribution $q(\theta)$ over an NN with weight matrices $W_i \in R^{K_i \times K_{i-1}}$ and bias vectors $b_i \in R^{K_i}$ for each layer by

$$\begin{aligned} &W_i = M_i \cdot \text{diag}\left(\left[z_{i,j}\right]_{j=1}^{K_i}\right) \\ &z_{i,j} \sim \text{ Bernoulli } (p_i) \text{ for } i = 1, \ldots, L, j = 1, \ldots, K_{i-1} \end{aligned} \tag{11}$$

where probabilities p_i and M_i are variational parameters and the binary variable $z_{i,j} = 0$ corresponding to the unit j in layer $i-1$ are dropped as input to layer i. The minimisation of the variational objective becomes

$$\mathcal{L}_{\text{MCDO}} := \frac{1}{N} \sum_{i=1}^{N} \|y_i - \hat{y}_i\|_2^2 + \lambda \sum_{i=1}^{L} \left(\|W_i\|_2^2 + \|b_i\|_2^2\right). \tag{12}$$

MCDO has been interpreted as VI [56]. Although the MCDO objective is not strictly an ELBO [131], we do sometimes refer to it as such. Analogously, other stochastic regularisation techniques [132, 133, 134] can also be reinterpreted as VI. Schlemper et al. [135] explore the applicability of MCDO to architectures, which are commonly used in medical image synthesis to model uncertainty for accelerated MR reconstructions. More generally, the majority of the works in medical image synthesis uses MCDO to approximate predictive uncertainty [135, 53, 71]. Indeed, MCDO is one of the most popular approximate inference schemes for complex deep learning models like CNNs, or Recurrent Neural Networks (RNNs) [136, 137]. Nonetheless, despite the impressive progress of BNN techniques, these technologies remain severely under-explored within medical image synthesis.

How to Measure Predictive Uncertainty?

Eq. (2) gives the mechanism to synthesise medical images and represents the full information of uncertainty of the imaging task. Here we differentiate metrics that summarise predictive uncertainty. The total uncertainty of the posterior predictive distribution $p(\hat{y}_q|x_q, \mathcal{D})$ is commonly measured by its variance $\mathbf{V}[\hat{y}_q|x_q]$. See Fig. 4 for results of a CNN model for diffusion MRI, which show the predictions of mean diffusivity (MD) and fractional anisotropy (FA), and their associated predictive uncertainty maps. The figure displays high correspondence between the root mean squared error (RMSE) maps and the predictive uncertainty on both FA and MD of a test subject, demonstrating the utility of the uncertainty map as a surrogate of prediction accuracy. It also shows strong correlation between the intensity value of the prediction and the predictive uncertainty, being in agreement with the observation that the error map itself correlates strongly with the intensity values.

To elucidate the sources of uncertainty, the total uncertainty can be further decomposed using the law of total variance:

$$\mathbf{V}[\hat{y}_q|x_q] = \underbrace{\mathbf{V}_{q^*(\theta)}\left[\mathbf{E}(\hat{y}_q|x_q, \theta)\right]}_{\Delta_{\text{E}}[\hat{y}_q]} + \underbrace{\mathbf{E}_{q^*(\theta)}\left[\mathbf{V}(\hat{y}_q|x_q, \theta)\right]}_{\Delta_{\text{A}}[\hat{y}_q]}, \tag{13}$$

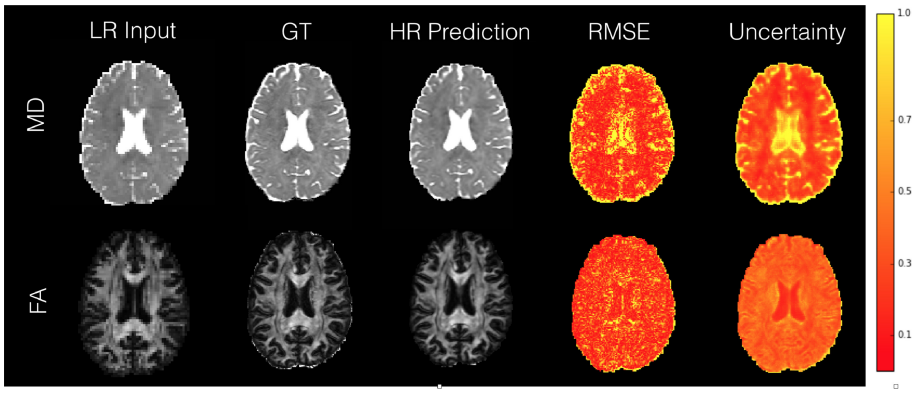


Figure 4: Comparison between voxel-wise RMSE and predictive uncertainty maps for FA and MD computed on a HCP test subject (min-max normalised for MD and FA separately). Low resolution input, ground truth and the mean of high resolution predictions are also shown. Source: [66].

where $\mathbf{E}(\hat{y}_q|x_q,\theta)$ and $\mathbf{V}(\hat{y}_q|x_q,\theta)$ are, respectively, the mean and variance of the prediction $\hat{y}_q$ according to $p(\hat{y}_q|x_q,\theta)$. The first term $\Delta_{\mathrm{E}}[\hat{y}_q]$ measures epistemic uncertainty since it ignores any contribution to the variance of $\hat{y}_q$ from the stochasticity in the data x_q. In contrast, the second term $\Delta_{\mathrm{A}}[\hat{y}_q]$ represents the average value of $\mathbf{V}(\hat{y}_q|x_q,\theta)$. This term ignores any contribution to the variance of $\hat{y}_q$ from θ and thus models aleatoric uncertainty. The importance of distinguishing between different forms of uncertainty has recently been recognised in deep learning models [57, 53]. We describe one approach in this direction by decomposing the predictive variance into aleatoric and epistemic components. The epistemic uncertainty (of parametric nature) can be obtained using BNNs and approximate inference schemes (e.g. VI or MCDO) thus it is encapsulated in the approximate posterior distribution. Meanwhile, quantifying aleatoric uncertainty can be captured by computing the variance of the likelihood. This broad class of models, where the variance is a function of the input, is often termed as input-dependent or heteroscedastic noise models [138, 139]. In practice, recent works rely on doubling the network architecture and modelling the likelihood as a Gaussian distribution with input-dependent varying variance (see Fig. 5), that is, $p(\hat{y}_q|x_q,\theta) = \mathcal{N}\left(\hat{y}; \mathrm{F}^{\mu}_{\theta_1}(x), \mathrm{F}^{\sigma}_{\theta_2}(x)\right)$, where $\mathrm{F}^{\mu}_{\theta_1}(\cdot)$ and $\mathrm{F}^{\sigma}_{\theta_2}(\cdot)$ refer to the "mean" and "covariance" networks respectively, with the approximate posterior distribution being $q_\psi(\theta = \{(\theta_1, \theta_2)\})$. Note that predictive uncertainty can be also decomposed by using homoscedastic noise models (i.e. constant variance across all spatial locations), but this approximation is highly unrealistic in medical image synthesis.

One can estimate the variance of a quantity of interest derived from a synthesised image and potentially decompose it into aleatoric and epistemic components. Let $f(\cdot)$ be any reasonably behaved function, which transforms the synthesised image $\hat{y}_q$ into a quantity of interest, and we estimate the variance

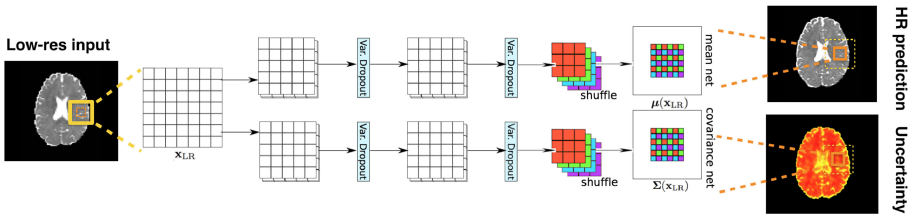


Figure 5: Illustration of a heteroscedastic network with variational dropout, with diagonal covariance. The top 3D-ESPCN estimates the mean and the bottom one estimates the covariance matrix of the likelihood. Variational dropout is applied to feature maps after every convolution, where Gaussian noise is injected into feature maps $F_{\text{out}} = \mu_Y + \epsilon \odot \sigma_Y$, with $\epsilon \sim \mathcal{N}(0, I)$. Source: [66].

in the transformed domain (i.e. $\mathbf{V}[f(\hat{y}_q)|x_q]$). If $f(\cdot)$ is an identity map, that is, $f(\hat{y}_q) = \hat{y}_q$, Eq. (13) can be approximated using T MC samples:

$$\widehat{\mathbf{V}}[\hat{y}_q|x_q] = \underbrace{\frac{1}{T}\sum_{t=1}^{T} \mathrm{F}^{\mu}_{\theta_1^t}(x_q)\mathrm{F}^{\mu}_{\theta_1^t}(x_q)^\top - \bar{\mathrm{F}}^{\mu}(x_q)\bar{\mathrm{F}}^{\mu}(x_q)^\top}_{\widehat{\Delta}_{\mathrm{E}}(\hat{y}_q)} + \underbrace{\sum_{i=1}^{T} \mathrm{F}^{\sigma}_{\theta_2^t}(x_q)}_{\widehat{\Delta}_{\mathrm{A}}(\hat{y}_q)}, \tag{14}$$

where $\bar{\mathrm{F}}^{\mu}(x_q) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{F}^{\mu}_{\theta_1^t}(x_q)$ with $\{(\theta_1^t, \theta_2^t)\}_{t=1}^{T} \sim q^*_\psi(\theta)$. If $f(\cdot)$ is "complicated" again we need to resort to MC sampling. Following Tanno et al. [53], given $\{(\theta_1^t, \theta_2^t)\}_{t=1}^{T} \sim q^*_\psi(\theta)$ and $\{f^t\}_{j=1}^{J} \sim p(\hat{y}_q|x_q, \theta_1^t, \theta_2^t)$ we estimate the propagated epistemic uncertainty $\Delta_{\mathrm{E}}[f(\hat{y}_q)]$ and propagated aleatoric $\Delta_{\mathrm{A}}[f(\hat{y}_q)]$ uncertainty as

$$\widehat{\Delta}_{\mathrm{E}}\left[f(\hat{y}_q)\right] := \frac{1}{T}\sum_{t}(\bar{f}^t)^2 - \left(\frac{1}{(J-1)T}\sum_{j,t} f_j^t\right)^2, \tag{15}$$

$$\widehat{\Delta}_{\mathrm{A}}\left[f(\hat{y}_q)\right] := \frac{1}{(J-1)T}\sum_{j,t}(f_j^t)^2 - \frac{1}{T}\sum_{t}(\bar{f}^t)^2, \tag{16}$$

$$\bar{f}^t := \frac{1}{J}\sum_{j} f_j^t. \tag{17}$$

Due to "double sampling", these estimators tend to have higher variance than the case where $f(\hat{y}_q) = \hat{y}_q$.

Instead of the variance of the posterior predictive distribution, we can also use its entropy as a measure of the overall predictive uncertainty [140]. The total uncertainty of the predictive distribution in Eq. (2) can then be quantified as $\mathbf{H}(\hat{y}_q|x_q)$, where $\mathbf{H}(\cdot)$ denotes the differential entropy of a probability distribution. This also allows decomposing predictive uncertainty into the two forms of uncertainty. The expectation of $\mathbf{H}\left(\hat{y}_q|x_q, \theta\right)$ under $q^*_\psi(\theta)$, that is, $\mathbf{E}_{q^*_\psi(\theta)}[\mathbf{H}(\hat{y}_q|x_q, \theta)]$, can be used to measure aleatoric uncertainty, and the

difference between total and aleatoric uncertainty to quantify the epistemic uncertainty:

$$\mathbf{H}[\hat{y}_q|x_q] - \mathbf{E}_{q^*_\psi(\theta)}[\mathbf{H}(\hat{y}_q|x_q,\theta)] := \mathbf{MI}(\hat{y}_q,\theta)\,, \tag{18}$$

which is the mutual information [141] between the posterior distributions of the model parameters θ and $\hat{y}_q$.

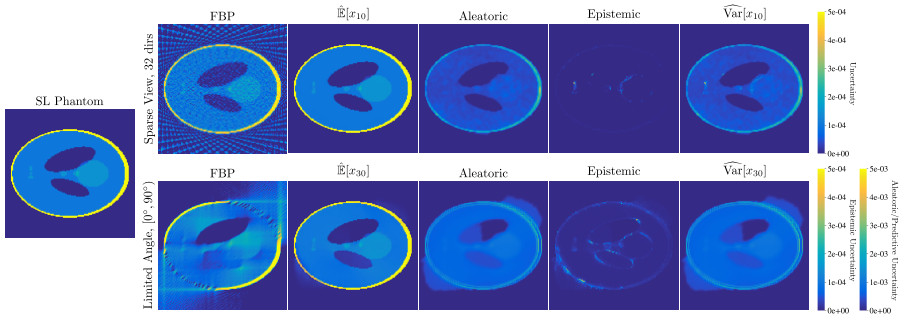


Figure 6: The reconstructions for sparse view CT with 32 directions (top) and limited angle with [0, 90°) (bottom). Source [130].

This decomposition allows us to separately quantify aleatoric and epistemic uncertainties. We give an illustration in Fig. 6 for CT reconstructions [130]. It is observed that in both sparse view and limited angle CT reconstructions, aleatoric uncertainty appears to dominate, with its overall shape close to the mean (but of a smaller magnitude). Epistemic uncertainty is localised to certain regions (and is of a smaller magnitude), capturing artefacts due to limited angle data. Thus, aleatoric and epistemic uncertainties provide complementary information about the reconstructions, and might provide different insights into their reliability.

Miscellaneous Approximations

Recent deep inferential machinery may also hold potential for the synthesis community. These techniques also employ deep neural networks but often obtain the associated uncertainties differently from BNNs. Below we describe the most influential ones. Adler et al. [142] employ a modified conditional Wasserstein generative adversarial network [143] to generate a high-dose CT from low-dose counterparts. However, the approach was only evaluated on simplified settings. Denker et al. [144] use conditional invertible neural networks which are inferential machinery based on (conditional) normalising flow [145, 146]. Normalising flow allows learning expressive conditional densities by maximum likelihood estimation. The authors aim to learn a conditional density of images from noisy low-dose CT measurements based on training data obtained from high-dose reconstructions. Tonolini et al. [147] and Zhang et al. [148] concurrently use a conditional variational autoencoder framework [125] for solving Bayesian image

reconstruction problems. Zhang et al. [148] provide the theoretical underpinning for approximate posterior inference and demonstrations on Gaussian and Poisson image denoising. More recently, Tezcan et al. [149] propose a hybrid approach for under-sampled MRI reconstruction to overcome the curse-of-dimensionality. The authors introduce a low-dimensional latent space given the acquisition data in k-space modelled via a variational autoencoder, and then apply MCMC for the sampling. In a yet slightly different vein, more recently, Edupuganti et al. [150] propose an approach for uncertainty quantification via variational autoencoders, with uncertainty encoded in the low-dimensional latent variable, and consistency enforced by minimising a loss based on the Stein unbiased risk estimator, and demonstrate the approach on MRI reconstruction.

Finally, an alternative approach to uncertainty quantification is ensembling (i.e. Bootstrap posteriors), where the variance of the predictions of multiple networks (i.e. the ensemble) is used to quantify predictive uncertainty [151]. In a number of settings, deep ensembles are becoming the gold standard approach for obtaining an accurate and well-calibrated posterior predictive distribution [152, 153, 154]. Within the machine learning community, the idea that deep ensembles should be regarded as an approximate approach to Bayesian marginalisation, instead of a competing (non-Bayesian) method to Bayesian inference, is emerging [83]. Pearce et al. [155] argue that deep ensembles perform approximate Bayesian inference, and Gustafsson et al. [156] also mention that deep ensembles can be regarded as samples from an approximate posterior distribution. Ensemble methods are limited by their computational cost as multiple NNs need to be trained independently (using different network initialisations). Furthermore, ensembling NNs requires even more significant memory and computational overhead at training and test time. To overcome the computational bottleneck, Huang et al. [157] among others [80, 158, 159] propose faster methods which train ensembles by leveraging different parameter configurations obtained in one single stochastic gradient descent trajectory. However, these methods come at the cost of reduced predictive performance [153]. There has been growing interest in uncertainty quantification using deterministic NNs which quantify uncertainty in a single forward pass and therefore have a smaller memory footprint [160, 161, 162].

All these approaches hold great potential for medical image synthesis.

Useful GitHub Repositories

Training a Bayesian NN efficiently is highly non-trivial. The current practice in machine learning strongly encourages the sharing of relevant implementations, mostly via GitHub. Instead of listing all existing links on Bayesian NNs, we would like to mention a few GitHub repositories that provide PyTorch implementations of the approximate inference methods that we have discussed, along with useful Google Colaboratory (Colab) notebooks. We believe that it is preferable to suggest exemplary implementations that are currently available on GitHub, and that have been carefully vetted by many members of the machine community. In this regard, we highly recommend the following GitHub reposi-

tory `https://github.com/JavierAntoran/Bayesian-Neural-Networks`, which has been redacted by Javier Antorán, a PhD candidate in the Machine Learning Group at Cambridge University. We include this repository for its richness, as well as its excellent readability. The author also provides Colab notebooks, which can be easily run without any need for expensive hardware, and allow interested readers to better familiarise themselves with different models. We would also like to mention Kumar Shridhar's repository `https://github.com/kumar-shridhar/PyTorch-BayesianCNN`, which includes Bayesian convolutional layers.

4 Open Challenges

So far we have discussed machine learning tools, which are currently available to the synthesis community to include a new dimension in synthesis applications: uncertainty quantification. In this final section, instead, we would like to point out several outstanding technical and clinical challenges. For instance, we are often forced to opt for restrictive, yet computationally feasible descriptors of reality over more truthful but computationally infeasible ones. In §4.1 we discuss the implications of the approximations we employ, and identify several possible research opportunities. In §4.2 we briefly discuss the additional hurdles we face when deploying uncertainty quantification technologies within the complex structure of healthcare, and envision that risk needs to be quantified in the context in which clinical decisions are formulated.

4.1 Can We Trust Uncertainty?

As is with all kind of quantifications, one is naturally interested in assessing whether we can actually trust the obtained uncertainty estimates; even more so if several approximations are taken. So, can we trust uncertainty? To answer the question we first present the sources of "(in)accuracy", putting a major focus on BNNs. We then argue that a quantitative evaluation of the uncertainty estimates would address at least (i) how accurate the estimates are (with respect to the ground-truth posterior distribution); and (ii) how robust they would be with respect to data distribution shift.

Sources of (In)accuracy

Computational feasibility often imposes restrictive approximations, leading to approximate likelihood, prior and posterior distributions, and thereby resulting in inaccurate estimation of aleatoric or epistemic uncertainty. Likelihood misspecification arises when overly simplistic assumptions are adopted for either the forward map or the noise statistics. Due to the high-dimensionality of the output, the likelihood is often assumed to be a Gaussian distribution with a diagonal covariance matrix, which provides only pixel-wise marginal distributions, and thus is unable to capture multi-modality of the predictive distribution (i.e. the presence of multiple modes). Further, the diagonality assumes

that the output pixels are statistically independent given the input. Likewise, the prior distribution $p(\theta)$ is prone to misspecification. This is especially true for data-driven approaches, where the parameters θ in neural networks (possibly due to severe over-parameterisation) often lack a clear semantic meaning or physical interpretation. This has largely prohibited domain practitioners from constructing hand-crafted priors. Instead of the result of the attempt to capture the modeller's prior knowledge (which is hard to grasp), priors are usually chosen (or at least in part) to ease computation, and as a result, in neural networks, one often contents with simple priors (i.e. the standard Gaussian distribution). Inevitably, this alters an orthodox interpretation of the prior in Bayesian statistics.

Even if the likelihood and the prior were both faithfully constructed to capture the real-world physics, the posterior distribution $p(\theta|\mathcal{D})$ is often approximated by Gaussian distributions with diagonal covariance (sometimes with low-rank or diagonal assumption), to facilitate or enable the requisite computation. Undoubtedly, this is a restrictive assumption. Foong et al. [163, 164] study the quality of common approximate inference methods VI and MCDO in approximating the Bayesian predictive distribution. They shed interesting insight into the pathologies of these approximation schemes, which up to now remain poorly understood. The issue of calibrating uncertainty estimates remains a big open question for both approximate inference techniques and deep learning based approaches, and it is currently an active area of research within the deep learning community [165, 166].

Practical Shortcomings of Bayesian Neural Networks

Bayesian methods have the potential to fix the shortcomings of deep learning (e.g. over-fitting, robustness, detection of out-of-distribution samples). Yet currently BNNs are often impractical and rarely match the performance of standard methods [167]. The impracticability of such deep inferential machineries can be attributed to several factors including (i) implementation complexity: BNNs are fairly sensitive to hyperparameter selection and initialisation strategies, and the training process can be substantially more challenging [168]; (ii) computational cost: BNNs can take orders of magnitude longer to converge than standard (deterministic) NNs, or alternatively, deep ensemble models require simultaneous training of multiple networks; (iii) weak performance: BNNs rely on crude approximations to achieve scalability, which often result in limited or unreliable uncertainty estimates [164]. In fact, the approximating family (e.g. $\mathcal{Q}_{\mathrm{FFG}}$) may not contain good approximations to the posterior distribution, and even if it does, the method (e.g. stochastic VI) may not be able to find a good approximate posterior within the chosen family. Not surprisingly, BNNs are rarely employed by the medical imaging community due to their complex deployment, which tends to overshadow their theoretical advantages.

The machine learning community proposed several solutions that partially address some of the pitfalls: recent works have largely focused on scalable inference [119, 167, 169, 170, 171, 103, 81]. However, these have not yet been

picked up by the medical imaging community, arguably due to the lack of communication between the two. Undoubtedly, the primary goal of this review is to bridge two different communities to inform the imaging community of the recent exciting developments in the machine learning community.

When it comes to uncertainty quantification, medical image synthesis practitioners often have blindly resorted to simple (as less expressive) Bayesian methods (e.g. MCDO) without a second thought. The machine learning community has recently proposed several solutions, which may have the potential to scale up to truly high-dimensional data regimes, as commonly occurring in practical medical imaging applications. Clearly, we still face a scalability issue. One thus may argue that if many of the available methods (if not all!) are not yet applicable to high-dimensional medical imaging problems, it is then acceptable to resort to MCDO. On the contrary, we believe that it is still worth informing the medical image community of the existence of more "sophisticated" methods, even if those are not yet applicable to medical imaging problems. Addressing the lack of scalability would open a myriad of research opportunities, which the synthesis community should seize. For example, Tezcan et al. [149] propose a novel method, which reveals a mature understanding of the limitations of the current approaches in Bayesian deep learning. Overcoming those led to a novel reconstruction algorithm.

Benchmarking Uncertainty Estimates

The lack of realistic ground truths has greatly hindered the quantitative evaluation of the accuracy of uncertainty estimates. In practice, it is often highly desirable to validate the accuracy of the approximation via golden standard MCMC, which however is infeasible for many real-world synthesis applications, since the distribution of interest $p(y|x)$ (i.e. the underlying data distribution) is almost always unknown or the resulting posterior is simply too costly even for the most advanced MCMC sampling algorithm. Nonetheless, it may be still possible to validate the aleatoric uncertainty by handcrafting a test dataset where the "ground truth" intrinsic noise is known (e.g. passing a set of medical images through a known stochastic transformation). The validation of the parametric uncertainty is by no means less challenging as the target distribution of interest $p(\theta|\mathcal{D})$ (i.e. the posterior distribution over the parameters) is not accessible. However, controllable and realistic means to edit input images (e.g. adding pathological structures or structural perturbations) would enable systematically studying what kinds of "out-of-distribution" structures can be detected through the analysis of parametric uncertainty for different Bayesian approximation schemes to NNs. There have been various attempts to use distributional shift while bench-marking parametric uncertainty [151, 152].

The robustness under data shifts of the uncertainty estimate is as well under scrutiny [152]. Robustness is strictly related to how well-calibrated uncertainty estimates are under domain shifts — in various settings, the test data distribution tends to deviate from the training environment due to sample bias[9] and

[9]Sample bias is of epistemic nature and reflects the fact that the data we observed is only

non-stationarity, which detracts from performance. This unfortunately occurs to uncertainty estimates as well (i.e. non-calibrated as the distribution changes). Robustness under distributional shift (e.g. the presence of out-of-distribution inputs) is necessary for the safe deployment in clinical practice in which distributional shift is widely prevalent. Therefore, predictive uncertainty must be well-calibrated to allow us to quantitatively assess the risk of a possible degradation of the synthesis task while sounding out unknown ground. This is critical since we would like to use uncertainty as a defensive mechanism against failures.

Future work should investigate the benefits of using more complex likelihood models (e.g. the correlations between neighbouring pixels may further improve the reconstruction quality) such as mixture models [45], diversity losses [172, 173, 174] and more powerful density estimators [175, 176, 177] as well as more structured and expressive posterior approximations [178, 179]. Moreover, finding answers to the queries above would shed insight on the clinical validation of predictive uncertainty as a measure of practical utility.

4.2 How to Communicate Uncertainty to Clinicians?

Last we discuss the challenges with communicating uncertainty to clinicians, and risk-aware uncertainty quantification, where the risk is related to the degree to which the synthesis has to be faithful. These challenges motivate revisiting the development of uncertainty analysis and quantification technologies.

Ideally, the translation of uncertainty quantification technologies from the machine learning community into clinical practice should cause as little disruption as possible to existing clinical workflows. There are several possible ways to convey uncertainty to clinicians. The uncertainty can be either directly handed over to clinicians as visuals by means of pixel-wise reliability scores (e.g. error bars or voxel-wise predictive variance) or summarising image-wise reliability scores (e.g. overall probability). Conveying uncertainty through visuals via voxel-wise variance appears more disruptive to clinical practice than a single reliability score. Having a one-off score is very tempting, but how will we actually go about deriving a single score from voxel-wise reliability maps or directly estimating a single score while foregoing the full Bayesian framework? It should be nothing less than a score that expresses whether the synthetic image is usable or not for the given task. For example, in the context of CT reconstruction, we may wish the score to inform us of the probability that a certain pathology (e.g. tumour or lesion) is present in the synthesised image. This issue can potentially be systematically addressed within the framework of hypothesis testing. Alternatively, uncertainty could play only behind closed doors, either embedded as a background defensive mechanism, or propagated through a pipeline (e.g. a cascade of inferential tasks for downstream decision-making).

How to optimally propagate uncertainty quantification in downstream analysis remains an open question, and is expected to be highly application de-

in part representative of the ground truth data distribution. If we train our model in presence of sampling bias, it is highly likely that it would poorly generalise towards under-represented features.

pendent: different downstream tasks would require uncertainty information of different quality. Indeed, we argue that the uncertainty quantification procedures should take the specifics of the downstream application into account, and we advocate for "granular" risk management as the risk depends on the downstream application.

We take radiation treatment planning as an example to show granularity of the risk-aware decision in image synthesis. For instance, if we had to synthesise a CT, which is often used to guide how to position radiation beams to target a tumour while avoiding healthy areas, we would not mind if there were defects (or high unreliability) in regions outside of the reach of the photon beams. Furthermore, larger or smaller margins could be drawn around the target — which we may want to treat or avoid — based on the reliability of the image. Consider a scenario in which a diagnostic decision is made based on a synthesised image. In order to make such a diagnostic decision, we need to quantify how reliable the image is. Taking Cohen's caricature example [34] (see Figure 2), which shows how a deep learning based algorithm can "hallucinate" cancer. If the clinician is somehow not investigating the cancer itself, this image might still be useful. Meanwhile, if the downstream task were radiotherapy treatment planning for the cancer, it would be a clear red flag not to use the image. The ideal scenario would be to quantify risk based on the details of the application. However, risk-sensitive uncertainty quantification raises several technical and conceptual challenges about how to apply a threshold to uncertainty (or to define an admissible set) for risk management.

It remains unclear how to use predictive uncertainty appropriately so that we can quantify the risks in the space in which the clinical decisions are made. This remains a completely open question, yet we recognise the enormous importance of future works in this direction, while realising the full potential of uncertainty quantification technologies in clinical practice. These discussions also have significant implications for technology development (e.g. developing technologies that directly deliver uncertainty estimates for the clinical practice of interest) to optimise the computational expense.

5 Concluding remarks

In this chapter we have provided an up-to-date overview of uncertainty quantification for medical image synthesis, including image reconstruction. In recent years, uncertainty quantification has been hailed as a very promising strategy to address the outstanding challenge, i.e. the lack of robustness of many deep learning based techniques, and thus has received much attention. We have described basic concepts in uncertainty analysis (e.g. predictive, aleatoric and epistemic uncertainty) and the potential benefits of providing uncertainty information in image synthesis along with the usual point estimators.

Conceptually, uncertainty reasoning can be carried out elegantly within a Bayesian framework, where all relevant information is represented by probability distributions and different sources of information can be integrated by Bayes'

formula. Nonetheless, this poses enormous computational challenges, especially with the complex models, which have arisen in deep learning. We have discussed representative computational techniques, including classical approximate inference strategies (e.g. MCMC, Laplace approximation and variational inference) along with the more recent Bayesian neural networks and Monte Carlo dropout. We have also pointed out relevant links to open source implementations available on GitHub repositories and discussed how to quantify the sources of uncertainty.

Lastly, we discussed the technical and clinical challenges associated with uncertainty quantification. The technical ones are largely concerned with calibration of the obtained uncertainty estimates. The clinical ones instead involve how to communicate the uncertainty information without disrupting existing medical pipelines.

In sum, uncertainty quantification holds enormous potential for medical image synthesis. However, there remain many outstanding technical and clinical challenges that have to be overcome before these technologies can be routinely deployed in clinical practice. This calls for further research from both theoretical and applied perspectives. Big practical challenges include developing scalable inference techniques, which are as non-intrusive as possible to the current imaging pipelines and providing clinically interpretable metrics for conveying useful uncertainty information. Theoretically, it is important to establish relevant mathematical-statistical guarantees for existing and forthcoming computational techniques.

References

[1] J. E. Iglesias, E. Konukoglu, D. Zikic, B. Glocker, K. Van Leemput, and B. Fischl, "Is synthesizing mri contrast useful for inter-modality analysis?," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 631–638, Springer, 2013.

[2] D. H. Ye, D. Zikic, B. Glocker, A. Criminisi, and E. Konukoglu, "Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 606–613, Springer, 2013.

[3] N. Burgos, M. J. Cardoso, K. Thielemans, M. Modat, S. Pedemonte, J. Dickson, A. Barnes, R. Ahmed, C. J. Mahoney, J. M. Schott, *et al.*, "Attenuation correction synthesis for hybrid PET-MR scanners: application to brain studies," *IEEE Trans. Med. Imag.*, vol. 33, no. 12, pp. 2332–2341, 2014.

[4] M. J. Cardoso, C. H. Sudre, M. Modat, and S. Ourselin, "Template-based multimodal joint generative model of brain data," in *International Conference on Information Processing in Medical Imaging*, pp. 17–29, Springer, 2015.

[5] A. F. Frangi, S. A. Tsaftaris, and J. L. Prince, "Simulation and synthesis in medical imaging," *IEEE Trans. Med. Imag.*, vol. 37, no. 3, pp. 673–679, 2018.

[6] M. Dashti and A. M. Stuart, "The bayesian approach to inverse problems," *arXiv preprint arXiv:1302.6989*, 2013.

[7] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, "Solving inverse problems using data-driven models," *Acta Numer.*, vol. 28, pp. 1–174, 2019.

[8] H. Van Nguyen, K. Zhou, and R. Vemulapalli, "Cross-domain synthesis of medical images using efficient location-sensitive deep network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 677–684, Springer, 2015.

[9] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris, "Multimodal MR synthesis via modality-invariant latent representation," *IEEE Trans. Med. Imag.*, vol. 37, no. 3, pp. 803–814, 2017.

[10] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, "Image synthesis in multi-contrast MRI with conditional generative adversarial networks," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2375–2388, 2019.

[11] D. Nie, X. Cao, Y. Gao, L. Wang, and D. Shen, "Estimating CT image from MRI data using 3D fully convolutional networks," in *Deep Learning and Data Labeling for Medical Applications*, pp. 170–178, Springer, 2016.

[12] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 417–425, Springer, 2017.

[13] J. M. Wolterink, A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg, and I. Išgum, "Deep MR to CT synthesis using unpaired data," in *International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 14–23, Springer, 2017.

[14] A. Ben-Cohen, E. Klang, S. P. Raskin, M. M. Amitai, and H. Greenspan, "Virtual PET images from CT data using deep convolutional networks: initial results," in *International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 49–57, Springer, 2017.

[15] L. Bi, J. Kim, A. Kumar, D. Feng, and M. Fulham, "Synthesis of positron emission tomography (PET) images via multi-channel generative adversarial networks (GANs)," in *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*, pp. 43–51, Springer, 2017.

[16] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "MedGAN: Medical image translation using GANs," *Comput. Med. Imag. Graphics*, vol. 79, p. 101684, 2020.

[17] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, *et al.*, "Dagan: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1310–1321, 2017.

[18] T. M. Quan, T. Nguyen-Duc, and W.-K. Jeong, "Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1488–1497, 2018.

[19] Y. Zhang, G. Wu, P.-T. Yap, Q. Feng, J. Lian, W. Chen, and D. Shen, "Hierarchical patch-based sparse representation—a new approach for resolution enhancement of 4D-CT lung data," *IEEE Trans. Med. Imag.*, vol. 31, no. 11, pp. 1993–2005, 2012.

[20] Y. Huang, L. Shao, and A. F. Frangi, "Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6070–6079, 2017.

[21] A. S. Chaudhari, Z. Fang, F. Kogan, J. Wood, K. J. Stevens, E. K. Gibbons, J. H. Lee, G. E. Gold, and B. A. Hargreaves, "Super-resolution musculoskeletal MRI using deep learning," *Magn. Reson. Med.*, vol. 80, no. 5, pp. 2139–2154, 2018.

[22] K. Sommer, A. Saalbach, T. Brosch, C. Hall, N. Cross, and J. Andre, "Correction of motion artifacts using a multiscale fully convolutional neural network," *Amer. J. Neurorad.*, vol. 41, no. 3, pp. 416–423, 2020.

[23] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pp. 241–246, IEEE, 2016.

[24] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in Neural Information Processing Systems*, pp. 341–349, 2012.

[25] F. Natterer and F. Wübbeling, *Mathematical Methods in Image Reconstruction*. SIAM, 2001.

[26] G. Ongie, A. Jalal, C. A. Baraniuk, R. G. Metzler, A. G. Dimakis, and R. Willett, "Deep learning techniques for inverse problems in imaging," *IEEE J. Sel. Areas in Inf. Theory*, vol. 1, no. 1, pp. 39–56, 2020.

[27] G. Wang, J. C. Ye, and B. De Man, "Deep learning for tomographic image reconstruction," *Nature Mach. Intel.*, vol. 2, pp. 737–748, 2020.

[28] P. Putzky and M. Welling, "Recurrent inference machines for solving inverse problems," *arXiv:1706.04008*, 2017.

[29] J. Adler and O. Öktem, "Learned primal-dual reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1322–1332, 2018.

[30] A. Hauptmann, F. Lucka, M. Betcke, N. Huynh, J. Adler, B. Cox, P. Beard, S. Ourselin, and S. Arridge, "Model-based learning for accelerated, limited-view 3-D photoacoustic tomography," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1382–1393, 2018.

[31] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing." arXiv:1912.10557, 2019.

[32] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, no. 7697, pp. 487–492, 2018.

[33] J. He, Y. Wang, and J. Ma, "Radon inversion via deep learning," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 2076–2087, 2020.

[34] J. P. Cohen, M. Luck, and S. Honari, "Distribution matching losses can hallucinate features in medical image translation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 529–536, Springer, 2018.

[35] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen, "On instabilities of deep learning in image reconstruction-does ai come at a cost?," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, no. 48, pp. 30088–30095, 2020.

[36] Y. Bengio, I. Goodfellow, and A. Courville, *Deep learning*, vol. 1. MIT press Massachusetts, USA:, 2017.

[37] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated mri data," *Magnetic Resonance in Medicine*, vol. 79, no. 6, pp. 3055–3071, 2018.

[38] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: nonlinear phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[39] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. theory*, vol. 52, no. 2, pp. 489–509, 2006.

[40] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[41] E. Begoli, T. Bhattacharya, and D. Kusnezov, "The need for uncertainty quantification in machine-assisted medical decision making," *Nature Machine Intelligence*, vol. 1, no. 1, pp. 20–23, 2019.

[42] T. J. Sullivan, *Introduction to Uncertainty Quantification*. Springer, 2015.

[43] H. Greenspan, R. Tanno, M. Erdt, T. Arbel, C. Baumgartner, A. Dalca, C. H. Sudre, W. M. Wells, K. Drechsler, M. G. Linguraru, *et al.*, *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures: First International Workshop, UNSURE 2019, and 8th International Workshop, CLIP 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings*, vol. 11840. Springer Nature, 2019.

[44] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: techniques, applications and challenges," in *arXiv:2011.06225*, 2020.

[45] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. A. Eslami, D. J. Rezende, and O. Ronneberger, "A probabilistic U-net for segmentation of ambiguous images," in *Advances in Neural Information Processing Systems*, pp. 6965–6975, 2018.

[46] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Med. Imag. Anal.*, vol. 59, p. 101557, 2020.

[47] S. Hu, D. Worrall, S. Knegt, B. Veeling, H. Huisman, and M. Welling, "Supervised uncertainty quantification for segmentation with multiple annotations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 137–145, Springer, 2019.

[48] A. G. Roy, S. Conjeti, N. Navab, C. Wachinger, A. D. N. Initiative, *et al.*, "Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control," *NeuroImage*, vol. 195, pp. 11–22, 2019.

[49] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," *Medical image analysis*, vol. 57, pp. 226–236, 2019.

[50] M. Jain, S. Lahlou, H. Nekoei, V. Butoi, P. Bertin, J. Rector-Brooks, M. Korablyov, and Y. Bengio, "DEUP: direct epistemic uncertainty prediction." Preprint, arXiv:2102.08501v1, 2021.

[51] S. C. Hora, "Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management," *Reliability Engineering & System Safety*, vol. 54, no. 2-3, pp. 217–223, 1996.

[52] B. M. Ayyub and G. J. Klir, *Uncertainty Modeling and Analysis in Engineering and the Sciences*. CRC Press, New York, 2006.

[53] R. Tanno, A. Ghosh, F. Grussu, E. Kaden, A. Criminisi, and D. C. Alexander, "Bayesian image quality transfer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 265–273, Springer, 2016.

[54] H. G. Matthies, "Quantifying uncertainty: modern computational representation of probability and applications," in *Extreme man-made and natural hazards in dynamics of structures*, pp. 105–135, Springer, 2007.

[55] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?," *Structural Safety*, vol. 31, no. 2, pp. 105–112, 2009.

[56] Y. Gal, *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

[57] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Advances in Neural Information Processing Systems*, pp. 5574–5584, 2017.

[58] S. Depeweg, *Modeling Epistemic and Aleatoric Uncertainty with Bayesian Neural Networks and Latent Variables*. PhD thesis, Technische Universität München, 2019.

[59] H. Wang, D. M. Levi, and S. A. Klein, "Intrinsic uncertainty and integration efficiency in bisection acuity," *Vision research*, vol. 36, no. 5, pp. 717–739, 1996.

[60] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction," *arXiv preprint arXiv:1910.09457*, 2019.

[61] U. Bhatt, Y. Zhang, J. Antorán, Q. V. Liao, P. Sattigeri, R. Fogliato, G. G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, *et al.*, "Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty," *arXiv preprint arXiv:2011.07586*, 2020.

[62] G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging*. Cambridge University Press, 2008.

[63] D. J. MacKay, *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992.

[64] O. O'Neill, "Linking trust to trustworthiness," *International Journal of Philosophical Studies*, vol. 26, no. 2, pp. 293–300, 2018.

[65] D. C. Alexander, D. Zikic, J. Zhang, H. Zhang, and A. Criminisi, "Image quality transfer via random forest regression: applications in diffusion MRI," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 225–232, Springer, 2014.

[66] R. Tanno, D. E. Worrall, E. Kaden, A. Ghosh, F. Grussu, A. Bizzi, S. N. Sotiropoulos, A. Criminisi, and D. C. Alexander, "Uncertainty modelling in deep learning for safer neuroimage enhancement: Demonstration in diffusion MRI," *NeuroImage*, p. 117366, 2020.

[67] J. C. Reinhold, Y. He, S. Han, Y. Chen, D. Gao, J. Lee, J. L. Prince, and A. Carass, "Validating uncertainty in medical image translation," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 95–98, IEEE, 2020.

[68] J. C. Reinhold, Y. He, S. Han, Y. Chen, D. Gao, J. Lee, J. L. Prince, and A. Carass, "Finding novelty with uncertainty," in *Medical Imaging 2020: Image Processing*, vol. 11313, p. 113130H, International Society for Optics and Photonics, 2020.

[69] K. Kläser, P. Borges, R. Shaw, M. Ranzini, M. Modat, D. Atkinson, K. Thielemans, B. Hutton, V. Goh, G. Cook, *et al.*, "Uncertainty-aware multi-resolution whole-body mr to ct synthesis," in *International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 110–119, Springer, 2020.

[70] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 655–663, Springer, 2018.

[71] R. Mehta, T. Christinck, T. Nair, P. Lemaitre, D. Arnold, and T. Arbel, "Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures*, pp. 23–32, Springer, 2019.

[72] D. J. MacKay, "A practical bayesian framework for backpropagation networks," *Neural computation*, vol. 4, no. 3, pp. 448–472, 1992.

[73] C. M. Bishop, "Bayesian methods for neural networks," tech. rep., Aston University, 1995.

[74] J. O. Berger and L. R. Pericchi, "The intrinsic bayes factor for model selection and prediction," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 109–122, 1996.

[75] J. O. Berger, *Statistical decision theory and Bayesian analysis*. Springer Series in Statistics, Springer-Verlag, New York, second ed., 1985.

[76] A. M. Stuart, "Inverse problems: a Bayesian perspective," *Acta Numer.*, vol. 19, pp. 451–559, 2010.

[77] K. Ito and B. Jin, *Inverse Problems: Tikhonov Theory and Algorithms*. World Scientific, Hackensack, NJ, 2015.

[78] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, 2015.

[79] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan, "Streaming variational bayes," *arXiv preprint arXiv:1307.6769*, 2013.

[80] T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, and A. G. Wilson, "Loss surfaces, mode connectivity, and fast ensembling of dnns," *arXiv preprint arXiv:1802.10026*, 2018.

[81] P. Izmailov, W. J. Maddox, P. Kirichenko, T. Garipov, D. Vetrov, and A. G. Wilson, "Subspace inference for bayesian deep learning," in *Uncertainty in Artificial Intelligence*, pp. 1169–1179, PMLR, 2020.

[82] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," *arXiv preprint arXiv:1803.05407*, 2018.

[83] A. G. Wilson and P. Izmailov, "Bayesian deep learning and a probabilistic perspective of generalization," *arXiv preprint arXiv:2002.08791*, 2020.

[84] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York, 2001.

[85] M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," *J. Royal Statist. Soc.: Ser. B*, vol. 73, no. 2, pp. 123–214, 2011.

[86] S. Pedemonte, C. Catana, and K. Van Leemput, "Bayesian tomographic reconstruction using Riemannian MCMC," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 619–626, Springer, 2015.

[87] I. S. Weir, "Fully bayesian reconstructions from single-photon emission computed tomography data," *J. Amer. Statist. Assoc.*, vol. 92, no. 437, pp. 49–60, 1997.

[88] É. Barat, C. Comtat, T. Dautremer, T. Montagu, and R. Trébossen, "A nonparametric Bayesian approach for PET reconstruction," in *2007 IEEE Nuclear Science Symposium Conference Record*, vol. 6, pp. 4155–4162, IEEE.

[89] M. Filipovic, E. Barat, T. Dautremer, C. Comtat, and S. Stute, "PET reconstruction of the posterior image probability, including multimodal images," *IEEE Trans. Med. Imag.*, vol. 38, no. 7, pp. 1643–1654, 2019.

[90] Q. Long, M. Scavino, R. Tempone, and S. Wang, "Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations," *Comput. Methods Appl. Mech. Engrg.*, vol. 259, pp. 24–39, 2013.

[91] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, pp. 183–233, 1999.

[92] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[93] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[94] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University of London, London, 2003.

[95] M. Opper and C. Archambeau, "The variational Gaussian approximation revisited," *Neural Comput.*, vol. 21, no. 3, pp. 786–792, 2009.

[96] E. Challis and D. Barber, "Gaussian Kullback-Leibler approximate inference," *J. Mach. Learn. Res.*, vol. 14, pp. 2239–2286, 2013.

[97] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *J. Mach. Learn. Res.*, vol. 14, pp. 1303–1347, 2013.

[98] T. P. Minka, *A family of algorithms for approximate Bayesian inference*. ProQuest LLC, Ann Arbor, MI, 2001. Thesis (Ph.D.)–Massachusetts Institute of Technology.

[99] C. Zhang, S. Arridge, and B. Jin, "Expectation propagation for Poisson data," *Inverse Problems*, vol. 35, no. 8, pp. 085006, 27, 2019.

[100] H. v. Rue, S. Martino, and N. Chopin, "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 71, no. 2, pp. 319–392, 2009.

[101] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statistics*, vol. 22, pp. 79–86, 1951.

[102] S. Mandt, M. D. Hoffman, and D. M. Blei, "Stochastic gradient descent as approximate bayesian inference," *J. Mach. Learn. Res.*, vol. 18, pp. 1–35, 2017.

[103] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson, "A simple baseline for bayesian uncertainty in deep learning," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, pp. 13153–13164, Curran Associates, Inc., 2019.

[104] Q. Liu and D. Wang, "Stein variational gradient descent: A general purpose bayesian inference algorithm," *Advances in Neural Information Processing Systems*, vol. 29, pp. 2378–2386, 2016.

[105] Q. Liu and D. Wang, "Stein variational gradient descent as moment matching," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, pp. 8854–8863, Curran Associates, Inc., 2018.

[106] Y. Wang and D. M. Blei, "Frequentist consistency of variational Bayes," *J. Amer. Statist. Assoc.*, vol. 114, no. 527, pp. 1147–1161, 2019.

[107] A. Repetti, M. Pereyra, and Y. Wiaux, "Scalable Bayesian uncertainty quantification in imaging inverse problems via convex optimization," *SIAM J. Imaging Sci.*, vol. 12, no. 1, pp. 87–118, 2019.

[108] A. G. Wilson, "The case for bayesian deep learning," *arXiv preprint arXiv:2001.10995*, 2020.

[109] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118. Springer Science & Business Media, 2012.

[110] D. J. MacKay, "Probable networks and plausible predictions—a review of practical bayesian methods for supervised neural networks," *Network: computation in neural systems*, vol. 6, no. 3, pp. 469–505, 1995.

[111] A. Graves, "Practical variational inference for neural networks," in *Advances in Neural Information Processing Systems*, pp. 2348–2356, 2011.

[112] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," *arXiv preprint arXiv:1505.05424*, 2015.

[113] J. M. Hernández-Lobato and R. Adams, "Probabilistic backpropagation for scalable learning of bayesian neural networks," in *International Conference on Machine Learning*, pp. 1861–1869, 2015.

[114] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*, pp. 1050–1059, 2016.

[115] J. S. Denker and Y. LeCun, "Transforming neural-net output levels to probability distributions," in *Proceedings of the 3rd International Conference on Neural Information Processing Systems*, pp. 853–859, 1990.

[116] Y. Li, J. M. Hernández-Lobato, and R. E. Turner, "Stochastic expectation propagation," *arXiv preprint arXiv:1506.04132*, 2015.

[117] J. Hernandez-Lobato, Y. Li, M. Rowland, T. Bui, D. Hernández-Lobato, and R. Turner, "Black-box alpha divergence minimization," in *International Conference on Machine Learning*, pp. 1511–1520, PMLR, 2016.

[118] Y. Li and R. E. Turner, "R\'enyi divergence variational inference," *arXiv preprint arXiv:1602.02311*, 2016.

[119] M. Khan, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal, and A. Srivastava, "Fast and scalable bayesian deep learning by weight-perturbation in adam," in *International Conference on Machine Learning*, pp. 2611–2620, PMLR, 2018.

[120] S. Sun, G. Zhang, J. Shi, and R. Grosse, "Functional variational bayesian neural networks," *arXiv preprint arXiv:1903.05779*, 2019.

[121] T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient Hamiltonian Monte Carlo," in *International Conference on Machine Learning*, pp. 1683–1691, 2014.

[122] Y.-A. Ma, T. Chen, and E. Fox, "A complete recipe for stochastic gradient mcmc," in *Advances in Neural Information Processing Systems*, pp. 2917–2925, 2015.

[123] R. M. Neal, "Bayesian learning via stochastic dynamics," in *Advances in Neural Information Processing Systems*, pp. 475–482, 1993.

[124] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th International Conference on Machine Learning*, pp. 681–688, 2011.

[125] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[126] C. Naesseth, F. Ruiz, S. Linderman, and D. Blei, "Reparameterization gradients through acceptance-rejection sampling algorithms," in *Artificial Intelligence and Statistics*, pp. 489–498, PMLR, 2017.

[127] F. R. Ruiz, M. T. R. AUEB, and D. Blei, "The generalized reparameterization gradient," in *Advances in Neural Information Processing Systems*, pp. 460–468, 2016.

[128] M. Figurnov, S. Mohamed, and A. Mnih, "Implicit reparameterization gradients," in *Advances in Neural Information Processing Systems*, pp. 441–452, 2018.

[129] R. Barbano, C. Zhang, S. Arridge, and B. Jin, "Quantifying model uncertainty in inverse problems via bayesian deep gradient descent," *arXiv preprint, arXiv:2007.09971.*

[130] R. Barbano, Ž. Kereta, C. Zhang, A. Hauptmann, S. Arridge, and B. Jin, "Quantifying sources of uncertainty in deep learning-based image reconstruction," *arXiv preprint arXiv:2011.08413*, 2020.

[131] J. Hron, A. Matthews, and Z. Ghahramani, "Variational bayesian dropout: pitfalls and fixes," in *International Conference on Machine Learning*, pp. 2019–2028, PMLR, 2018.

[132] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," *arXiv preprint arXiv:1506.02557*, 2015.

[133] M. Teye, H. Azizpour, and K. Smith, "Bayesian uncertainty estimation for batch normalized deep networks," in *International Conference on Machine Learning*, pp. 4907–4916, PMLR, 2018.

[134] Y. Wen, P. Vicol, J. Ba, D. Tran, and R. Grosse, "Flipout: Efficient pseudo-independent weight perturbations on mini-batches," *arXiv preprint arXiv:1803.04386*, 2018.

[135] J. Schlemper, D. C. Castro, W. Bai, C. Qin, O. Oktay, J. Duan, A. N. Price, J. Hajnal, and D. Rueckert, "Bayesian deep learning for accelerated mr image reconstruction," in *International Workshop on Machine Learning for Medical Image Reconstruction*, pp. 64–71, Springer, 2018.

[136] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," *arXiv preprint arXiv:1506.02158*, 2015.

[137] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Advances in Neural Information Processing Systems*, pp. 1019–1027, 2016.

[138] D. A. Nix and A. S. Weigend, "Estimating the mean and variance of the target probability distribution," in *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, vol. 1, pp. 55–60, IEEE, 1994.

[139] C. R. Rao, "Estimation of heteroscedastic variances in linear models," *Journal of the American Statistical Association*, vol. 65, no. 329, pp. 161–172, 1970.

[140] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning," in *International Conference on Machine Learning*, pp. 1184–1193, PMLR, 2018.

[141] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[142] J. Adler and O. Öktem, "Deep bayesian inversion," *arXiv preprint arXiv:1811.05910*, 2018.

[143] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875*, 2017.

[144] A. Denker, M. Schmidt, J. Leuschner, P. Maass, and J. Behrmann, "Conditional normalizing flows for low-dose computed tomography image reconstruction," *arXiv preprint arXiv:2006.06270*, 2020.

[145] C. Winkler, D. Worrall, E. Hoogeboom, and M. Welling, "Learning likelihoods with conditional normalizing flows," *arXiv preprint arXiv:1912.00042*, 2019.

[146] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *arXiv preprint arXiv:1912.02762*, 2019.

[147] F. Tonolini, J. Radford, A. Turpin, D. Faccio, and R. Murray-Smith, "Variational inference for computational imaging inverse problems," *J. Mach. Learn. Res.*, vol. 21, no. 179, pp. 1–46, 2020.

[148] C. Zhang and B. Jin, "Probabilistic residual learning for aleatoric uncertainty in image restoration," *arXiv preprint arXiv:1908.01010*, 2019.

[149] K. C. Tezcan, C. F. Baumgartner, and E. Konukoglu, "Sampling possible reconstructions of undersampled acquisitions in mr imaging," *arXiv preprint arXiv:2010.00042*, 2020.

[150] V. Edupuganti, M. Mardani, S. Vasanawala, and J. Pauly, "Uncertainty quantification in deep MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 239–250, 2021.

[151] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.

[152] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, pp. 13991–14002, 2019.

[153] A. Ashukha, A. Lyzhov, D. Molchanov, and D. Vetrov, "Pitfalls of in-domain uncertainty estimation and ensembling in deep learning," *arXiv preprint arXiv:2002.06470*, 2020.

[154] F. Wenzel, K. Roth, B. S. Veeling, J. Światkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin, "How good is the bayes posterior in deep neural networks really?," *arXiv preprint arXiv:2002.02405*, 2020.

[155] T. Pearce, F. Leibfried, and A. Brintrup, "Uncertainty in neural networks: Approximately bayesian ensembling," in *International conference on artificial intelligence and statistics*, pp. 234–244, PMLR, 2020.

[156] F. K. Gustafsson, M. Danelljan, and T. B. Schon, "Evaluating scalable bayesian deep learning methods for robust computer vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 318–319, 2020.

[157] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot ensembles: Train 1, get m for free," *arXiv preprint arXiv:1704.00109*, 2017.

[158] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson, "A simple baseline for bayesian uncertainty in deep learning," *Advances in Neural Information Processing Systems*, vol. 32, pp. 13153–13164, 2019.

[159] S. Fort and S. Jastrzebski, "Large scale structure of neural network loss landscapes," *arXiv preprint arXiv:1906.04724*, 2019.

[160] J. van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *International Conference on Machine Learning, PMLR*, pp. 9690–9700, 2020.

[161] J. Z. Liu, L. Lin, S. Padhy, D. Tran, T. Bedrax-Weiss, , and B. Lakshminarayanan, "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," in *NeurIPS*, 2020.

[162] J. Mukhoti, A. Kirsch, J. van Amersfoort, P. H. Torr, and Y. Gal, "Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty." Preprint, arXiv:2102.11582v1, 2021.

[163] A. Y. Foong, D. R. Burt, Y. Li, and R. E. Turner, "On the expressiveness of approximate inference in bayesian neural networks," *arXiv preprint arXiv:1909.00719*, 2019.

[164] A. Y. Foong, Y. Li, J. M. Hernández-Lobato, and R. E. Turner, "'in-between'uncertainty in bayesian neural networks," *arXiv preprint arXiv:1906.11537*, 2019.

[165] A. Kumar, P. Liang, and T. Ma, "Verified uncertainty calibration." NeurIPS 2019, arXiv:1909.10155, 2019.

[166] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. Torr, and P. K. Dokania, "Calibrating deep neural networks using focal loss." NeurIPS 2020, arXiv:2002.09437, 2020.

[167] K. Osawa, S. Swaroop, A. Jain, R. Eschenhagen, R. E. Turner, R. Yokota, and M. E. Khan, "Practical deep learning with Bayesian principles." Preprint, arXiv:1906.02506v2, 2019.

[168] S. Rossi, P. Michiardi, and M. Filippone, "Good initializations of variational bayes for deep models," in *International Conference on Machine Learning*, pp. 5487–5497, PMLR, 2019.

[169] S. Farquhar, M. A. Osborne, and Y. Gal, "Radial bayesian neural networks: Beyond discrete support in large-scale bayesian deep learning," in *International Conference on Artificial Intelligence and Statistics*, pp. 1352–1362, PMLR, 2020.

[170] E. Daxberger, E. Nalisnick, J. U. Allingham, J. Antorán, and J. M. Hernández-Lobato, "Expressive yet tractable bayesian deep learning via subnetwork inference," *arXiv preprint arXiv:2010.14689*, 2020.

[171] J. Antorán, J. U. Allingham, and J. M. Hernández-Lobato, "Depth uncertainty in neural networks," *arXiv preprint arXiv:2006.08437*, 2020.

[172] D. Bouchacourt, P. K. Mudigonda, and S. Nowozin, "Disco nets: Dissimilarity coefficients networks," in *Advances in Neural Information Processing Systems*, pp. 352–360, 2016.

[173] A. Guzman-Rivera, D. Batra, and P. Kohli, "Multiple choice learning: Learning to produce multiple structured outputs," *Advances in neural information processing systems*, vol. 25, pp. 1799–1807, 2012.

[174] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 35–51, 2018.

[175] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172–189, 2018.

[176] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International Conference on Machine Learning*, pp. 2642–2651, PMLR, 2017.

[177] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," *arXiv preprint arXiv:1505.05770*, 2015.

[178] C. Louizos and M. Welling, "Structured and efficient variational deep learning with matrix gaussian posteriors," in *International Conference on Machine Learning*, pp. 1708–1716, 2016.

[179] M. D. Hoffman and D. M. Blei, "Structured stochastic variational inference," in *Artificial Intelligence and Statistics*, 2015.