Position Paper: Uncertainty Quantification in Deep Learning Is Unsatisfactory for Clinical Applications and Complex Decision Making

Ciaran Bench
Department of Data Science and AI
National Physical Laboratory
Teddington, UK
ciaran.bench@npl.co.uk

***Abstract*—Deep learning has significant potential to enhance medical services, but low tolerance for error and the risk of poor performance on unseen data has slowed the widespread adoption of models in practical settings. In principle, uncertainty quantification (UQ) may be used to evaluate the trustworthiness of predictions, facilitating the effective use of models in medical applications. UQ techniques in deep learning aim to reliably express the doubt in a measurement/prediction. However, common UQ techniques and evaluation metrics/measures in deep learning only consider uncertainty reliability as viewed from relatively simple measurement frameworks where contextual factors relevant to complex medical decision making can not be easily integrated. But even for cases where these factors may be quantified and considered, common methods do not achieve/assess all aspects of reliability that are relevant to clinical applications. We describe these shortcomings, and propose research priorities to help improve the effectiveness of UQ for medical applications, and realise the positive impact deep learning could have on patient outcomes.**

***Index Terms*—uncertainty quantification, deep learning, uncertainty reliability**

I. Introduction

While deep learning holds considerable potential to enhance medical services and improve patient outcomes, the commercial uptake of models has seen relative stagnation compared to other fields [1]. Among several reasons (e.g. privacy/logistical challenges with acquiring suitable datasets), a low tolerance for error imposes strict requirements on model performance that can be difficult to achieve, or even define in practice. E.g., the black box nature of many deep learning approaches can make it challenging to understand all of the relevant ways a model may fail in practical settings; yet, this is necessary to define suitable performance criteria. This lack of transparency also makes it appealing to provide some indication of the *trustworthiness* of individual predictions to facilitate deployment in real measurement scenarios (e.g. determine whether a prediction is likely to be correct, and therefore whether it can be trusted to inform diagnosis). Despite the potential of various techniques that aim to provide this information (explainable AI [2], and causal model development [3]), their practical utility remains questionable for many use-cases [4], [5].

Here, we focus on uncertainty quantification (UQ) techniques in deep learning, which can be used to estimate the degree of doubt in a model's prediction. In principle, this can provide a means to determine when a model is likely to output an incorrect prediction without prior knowledge of the ground truth. While this does not provide rich information about *why* a model may fail (i.e. provide some notion of interpretability that is often desired in medical decision making), it can be used to make more effective use of model predictions (e.g. to disregard predictions likely to be inaccurate, or to enhance efforts to prototype models/collate datasets). With that said, even in this capacity, we argue in this paper that commonly used UQ techniques and measures for assessing the quality of uncertainty estimates are not sufficient for medical applications.

The potential uses of model predictions in medicine are broad; e.g. enhancing the efficiency of a device already routinely used by human clinicians to plan treatment/inform diagnosis (i.e. physical measurements), or as autonomous agents capable of making complex decisions related to diagnosis or treatment plans without human intervention [6]. Yet, here, we argue that the frameworks used to quantify uncertainty in deep learning models are generally unsatisfactory given the practical challenges imposed in medical contexts.

For autonomous complex decision making, deep learning architectures and UQ methods for deep learning do not factor in all relevant contextual factors (hermeneutic uncertainty) which lack an effective means of quantification, and therefore, integration into deep learning frameworks [7], [8]. Instead, UQ methods in deep learning are formulated to try and achieve *uncertainty reliability* [9] as viewed from a measurement scenario where the relevant contextual factors are quantifiable and more readily integrated into model optimisation frameworks. This makes current UQ methods more appealing for models intended to provide, replace, or enhance an existing measurement service (where relevant confounding factors can

The project (22HLT01 QUMPHY) has received funding from the European Partnership on Metrology, co-financed from the European Union's Horizon Europe Research and Innovation Programme and by the Participating States. Funding for NPL was provided by Innovate UK under the Horizon Europe Guarantee Extension, grant numbers 10084125.

be more easily quantified). This scenario reflects how non-network based measurement techniques are used by clinicians in standard work flows. But even here, limitations in common strategies for quantifying uncertainty, and assessing uncertainty reliability in deep learning make it challenging to effectively implement and assess their practical utility given the conditions imposed in medical contexts. Therefore, we suggest that the effectiveness of UQ techniques in medical applications remains debatable even in more simple measurement scenarios.

A. Aim

Here we describe the inadequacies of common UQ techniques and UQ evaluation measures in deep learning with respect to the challenges involved with using models in clinical applications. We also include some discussion on the prospect of using UQ in generative deep learning, and propose tractable next steps for improving UQ techniques in a way that can improve clinical decision making. While the literature often refers to uncertainty reliability measures as 'metrics', this has specific implications in some fields that may not apply; instead we use the term 'measure' where applicable [10].

II. Uncertainty quantification in deep learning

While most deep learning models are deterministic, (i.e. evaluating the same input several times will produce identical point predictions), uncertainty quantification techniques in deep learning aim to move beyond point predictions and configure models so they output an uncertainty representation alongside predictions (e.g. a predictive distribution, a calibrated interval, or a set of plausible labels). While generic classification models already predict a distribution, these are often overconfident [11] and only capture one source of uncertainty (aleatoric uncertainty). Of several sources of uncertainty, epistemic uncertainty (reducible model uncertainty), and aleatoric uncertainty (irreducible uncertainty intrinsic to the data/task) are the most relevant in practical measurement scenarios. Hermeneutic uncertainty is relevant to determining the effective use of model outputs in complex decision making, but is not easily factored into modelling approaches given the lack of effective means for quantification [7], [8].

There are several different techniques for uncertainty quantification in deep learning (e.g. post-hoc recalibration, intrinsic modelling, and post-hoc ensemble summarised in Table I), each of which employ different theoretical frameworks for estimating various sources of uncertainty [12]. Frequentist methods treat uncertainty as the long-run frequency of events. E.g. temperature scaling (a post-hoc recalibration method) uses a hold-out calibration set to rescale class probabilities so they are calibrated against the observed distribution of prediction errors [11]. A scalar temperature parameter $T > 0$ is applied to the logits output by the model. Given a model's uncalibrated logits $\mathbf{z}$, where $\mathbf{z} = [z_1, z_2, \ldots, z_C]$ where C is the number of classes, the calibrated probabilities are given by:

$$p_i = \frac{\exp(z_i/T)}{\sum_{j=1}^{C} \exp(z_j/T)}, \quad i = 1, \ldots, C \tag{1}$$

where p_i is the probability of class i after temperature scaling. Conformal prediction is another post-hoc recalibration approach where (in the case of split conformal prediction [13]) nonconformity scores (e.g. residuals) of a hold-out calibration set (n examples) are used to define a conformal quantile for a chosen coverage level by taking the $(1-\alpha)$ quantile of the sorted scores (i.e. the k smallest score, where $k = \lceil (n+1)(1-\alpha) \rceil$). The prediction intervals for test examples are then derived using this quantile. However, these post-hoc recalibration methods may yield unreliable estimates with limited data [14].

Bayesian approaches, by contrast, incorporate prior beliefs that are refined with new evidence to produce predictive distributions [15]. Bayesian-inspired techniques are reasoned from the principles underlying Bayesian inference, which is a common theoretical framework for quantifying the uncertainty in a predictive model (epistemic uncertainty).

In principle, Bayesian inference can be used to acquire a distribution of predictions. However, with deep networks, the denominator (evidence) makes the posterior intractable to evaluate given the large number of parameters [16]. This motivates a need for more efficient techniques. Variational inference is one approach [16], where a more tractable variational distribution is used to approximate the posterior by optimising the evidence lower bound (ELBO). However, even variational inference is not tractable for large models [16], and there have been further efforts to develop even more efficient and effective techniques such as Monte Carlo Dropout or the Improved Variational Online Newton [17] that approximate variational inference (both intrinsic modelling approaches).

Evidential deep learning is a framework based on subjective logic theory where a model is trained to output evidence that is used to define a Dirichlet distribution over the class probabilities (i.e. a distribution over the possible combinations of output class probabilities) [18]. I.e. the model outputs the amount of evidence supporting the statement that the input belongs to a particular class. This evidence may be interpreted as a pseudo-count of training examples that take the value of a class that the test input has resemblance to. A belief mass and uncertainty mass can be calculated for a given prediction. For the regression case, a Normal Inverse-Gamma distribution can be used instead.

Deep ensembles (post-hoc ensemble approach) stands apart from the other approaches as a discrete predictive distribution as derived by aggregating outputs from multiple identical models, each trained independently with distinct random weight initialisations [19].

Aleatoric uncertainty is often modelled by parameterising the model to predict the parameters of a distribution (e.g. a Gaussian Negative Log likelihood) in the case of regression, or by parameterising classification models to predict the variance on the logits [20].

III. Evaluating uncertainty reliability

There are several compelling use cases for AI models where the model is intended to perform a straightforward measure-

TABLE I
COMMON UNCERTAINTY QUANTIFICATION METHODOLOGIES IN DEEP LEARNING WITH DESCRIPTIONS (ADAPTED FROM [12])

UQ Methodology		Description	UQ Techniques
Post-hoc	Post-hoc ensemble	Uncertainties are derived by aggregating the outputs of multiple versions of the same architecture that are trained to perform the same task.	Deep Ensembles
	Post-hoc recalibration	A held-out calibration set is used to learn a transformation applied to the predictions to improve their calibration.	Temperature Scaling, Conformal Prediction, Isotonic Regression
Intrinsic		Uncertainties are estimated as a consequence of the design choices involved with the model's optimisation/evaluation.	Monte-Carlo Dropout, Quantile Regression, Evidential deep learning

ment (e.g. of a physiological parameter, or diagnosis) with clear notions of prediction error. Akin to how an uncertainty from a non-network based medical device may be used to assess the trustworthiness of a measurement, an uncertainty indicating the accuracy of a model's prediction may help improve the use of predictions in some clinical work flow (e.g. reject measurements/predictions likely to be inaccurate). Common deep learning frameworks, UQ techniques in deep learning, and measures used to evaluate the quality of uncertainty estimates cater to these straightforward measurement scenarios where relevant confounding factors/sources of uncertainty are readily quantifiable; here, we provide an overview of the key concepts related to evaluating the quality of estimated uncertainties from this perspective.

Two important concepts related to uncertainty reliability are calibration (i.e. the extent to which the magnitude of uncertainties correlate with the magnitude of prediction error) and sharpness (how tightly the predictive distribution falls around the ground truth) [21]. These can be assessed at various *scales*. E.g. global calibration measures assess whether the average uncertainty over the test set of examples is equivalent to the average prediction error. Local calibration measures, on the other hand, assess the equivalence of the average uncertainty and the prediction error for test predictions binned by uncertainty magnitude. This concept can be extended to the level of individual predictions, i.e. individual calibration/reliability [9]. One can also bin predictions by something other than uncertainty magnitude, e.g. by target class or data quality parameters to assess adaptivity [9]. Reliability diagrams that plot the extent to which each bin is calibrated can show local trends in over or underestimation. Here, there is an inherent tradeoff between locality and the effectiveness of the assessment of uncertainty reliability; i.e. some calibration/reliability measures increase with bin number.

The chosen expression of uncertainty (e.g. class probability or entropy for classification), and choice of reliability measure can influence perceptions of uncertainty reliability [12], [22]. Broadly there are two types of evaluation measures for regression: coverage and variance based measures. The former assesses whether ground truths fall within the X% confidence interval X% of the time, while variance-based measures assess the equivalence of uncertainty magnitude with prediction error. However, coverage based measures have been shown to lack suitable properties as uncertainty reliability measures in some cases [23].

The concepts presented here cater to relatively simple measurement scenarios. They only consider whether the uncertainty magnitude indicates the accuracy of the predicted measurand (quantity being predicted); i.e. they are agnostic to the effect an uncertainty estimate may have on patient outcomes given how it may be used within a given clinical work flow (this is a general weakness that also applies to the non-network based predictions). Therefore, this framework for evaluating uncertainty reliability does not necessarily indicate the practical utility of uncertainty estimates [24], [25]. This framework for uncertainty reliability also does not readily apply to some common applications of deep learning, e.g. reliability metrics/measures are challenging to apply to cases where there are less clear notions of prediction error (e.g. generative modelling). These limitations impose significant barriers to using UQ to assess the trustworthiness of AI systems used in medicine.

IV. INADEQUACIES OF UQ METHODS AND UNCERTAINTY EVALUATION MEASURES

We provide commentary on whether UQ techniques and reliability measures in deep learning provide the necessary insight as to whether the use of estimated uncertainties may improve patient outcomes given how model outputs may be used in practice.

A. Accounting for individualised contextual information

The methodological limitations of deep learning models make it challenging to incorporate the individualised assessment of various contextual factors that medical professionals sometimes use to influence decision making [7], [8]. These factors are particularly important for models intended to take the place of humans (partially or fully). For example, in the hypothetical case where an AI model is trained to observe symptom patterns to recommend treatment (as discussed in [7]), the disclosure/reporting of a patient's predicted condition can actually lead to negative patient outcomes in some cases, e.g. emerging due to domestic abuse. Similarly, a risk assessment for data-driven atrial fibrillation (period of irregular uncoordinated heartbeat, abbreviated as AF) classification [26] emphasises the role patient information (e.g. metadata, or

other symptoms) not accounted for by the model may play in improving patient outcomes. While medical professionals can infer whether this may be worth considering from contextual information acquired from repeated visits, deep learning models are not typically developed in a way to use this information. This is challenging to incorporate into a modelling framework given the highly individualised nature of the kind of context that may influence decision making, and also the unique way this can affect patient outcomes on a case-by-case basis. Effective frameworks for using current AI models [27] in medical contexts should therefore involve deliberation with medical professionals to define these factors with respect to existing ethical frameworks and attempt to formulate decision making criteria, or enhance model development in a way that takes these into account. Greater emphasis should be placed on incorporating contextual factors to formulate richer expressions of uncertainty and evaluations of reliability. This will become increasingly relevant as trends move towards models making more complex decisions.

With that said, in some cases, models may be used to automate procedural manual analysis of medical data, or to streamline/enhance an existing device used to make decisions; and in this capacity, popular UQ techniques in deep learning, in principle, may be more suitable, as relevant factors are more easily incorporated into their formulation. The rest of this article considers these use cases.

B. Individual reliability

In many applications, a single measurement or few measurements are used to inform decisions. Therefore, individual estimates of uncertainty should exhibit reliability (i.e. models should achieve individual reliability [28]). [29] provides one formal definition for *individual calibration*. Here, $\mathbf{H}$ (bold letters signify random variables in this section) is a randomised forecaster that maps an input x to a predictive cumulative distribution function (CDF). If $\mathbf{H}$ is a perfect forecaster then $\mathbf{H}[x]$ should be equal to the true CDF $F_{\mathbf{Y}|x}$, where the true outcome is a random variable $\mathbf{Y} \sim F_{\mathbf{Y}|x}$ (while $F_{\mathbf{Y}|x}$ is a CDF, here it is also used to represent the distribution of $\mathbf{Y}$ as in [29]). The probability integral transform $F_{\mathbf{Y}|x}(\mathbf{Y})$ should be a uniformly distributed variable if $\mathbf{Y}$ is indeed drawn from a distribution defined by $F_{\mathbf{Y}|x}$. So $\mathbf{H}[x](\mathbf{Y})$ (forecaster evaluated on the outcome) should behave like a variable sampled from a uniform distribution if $\mathbf{H}[x] = F_{\mathbf{Y}|x}$. Therefore, individual calibration $\mathrm{err}_{\mathbf{H}}(x)$ can be defined as the distance d between the CDF of $\mathbf{H}[x](\mathbf{Y})$ denoted as $F_{\mathbf{H}[x](\mathbf{Y})}$, and F_U, the CDF of a uniformly distributed random variable, giving $\mathrm{err}_{\mathbf{H}}(x) := d(F_{\mathbf{H}[x](\mathbf{Y})}, F_U)$.

However, many UQ techniques in deep learning do not optimise for individual reliability, and even those that do often fail to achieve it in practice or are not practically useful in medical contexts. Some generic implementations of post-hoc methods rescale uncertainty estimates based on global reliability measures, and therefore, do not encourage individual reliability [30]. Scalable Bayesian inspired approaches are not typically posed in a way that directly optimises for individual reliability, e.g., MCD approximates the minimisation of the ELBO, which does not optimise for individual reliability. Fig. 1 shows the results (bivariate histogram of predicted uncertainty and prediction error) of a deep learning model trained to predict blood pressure from raw photoplethysmography (PPG) signals, where MCD is used to estimate uncertainty. It is evident that MCD did not produce locally calibrated uncertainties. While the Gaussian negative log likelihood encodes individual reliability, in practice, this loss is not effective at achieving this. E.g. estimates given in [12] show that the model exhibits poor local reliability.

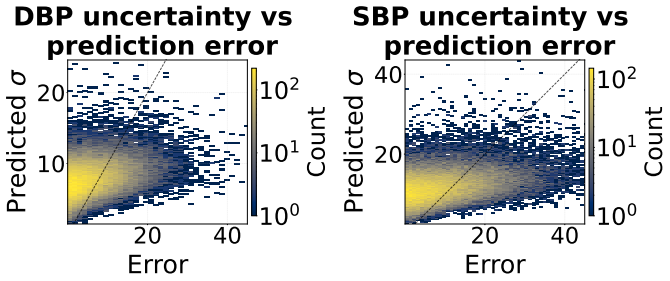

Fig. 1. Bivariate histograms showing the distribution of predicted uncertainties (combined epistemic and aleatoric modelled with a likelihood-based loss and MCD) with respect to prediction error for blood pressure (systolic and diastolic blood pressure, abbreviated SBP and DBP respectively) regression models trained on raw PPG time series. The dashed line indicates perfect calibration; here, the estimates clearly do not exhibit individual calibration. Figure created from data reported in [12].

Evidential deep learning encourages individual reliability (e.g. the KL divergence term of the loss pushes the predicted Dirichlet distribution away from the one-hot ground truth when the evidence higher than it should be), but is sensitive to class imbalance common in medical datasets [31]. Furthermore, the expression of uncertainty is not immediately interpretable like a class probability could be, which would require additional effort and validation to formulate an effective decision criteria. Some variants of conformal prediction may have the capacity to optimise for individual reliability [28], yet the need for a representative and large held-out calibration set imposes limitations for medical applications, along with the approach's other drawbacks. These also exhibit poor generalisability compared to intrinsic approaches [32]. Nevertheless, this perhaps is the most promising approach to take in cases where there are large datasets. Yet, there remains a need for UQ methods that optimise for individual reliability that also accommodate common constraints on available data (class imbalance, and general number of samples). While some variants of popular UQ techniques aim to address some of these limitations (e.g. [30], [31]), further work needs to be done to test their effectiveness at scale and in realistic contexts.

Furthermore, common measures for assessing uncertainty reliability in deep learning do not provide a quantitative indication of individual reliability. Local (expected calibration error ECE [11], uncertainty calibration error UCE [33], and variation calibration error VCE [12]) or global calibration measures (e.g. negative log likelihood, continuous ranked

probability score [34]) are more commonly implemented. Proper scoring rules like continuous ranked probability score [34] encode per instance sharpness and accuracy; ultimately they are used as aggregate global metrics, where this and the encoding of accuracy adds some ambiguity to the assessment of individual reliability.

C. Generalisability of uncertainty reliability

1) Adaptivity: Local reliability measures indicate the reliability of uncertainty estimates at various magnitudes, which can help facilitate decision making (e.g. high uncertainty predictions are reliable, and therefore can be used to reject predictions with confidence). However, it is also of interest to assess model performance on various data classes. E.g. are high uncertainties on noisier time series still reliable? Or for particular demographic profiles? Therefore it is crucial for models to produce uncertainty estimates that achieve adaptivity.

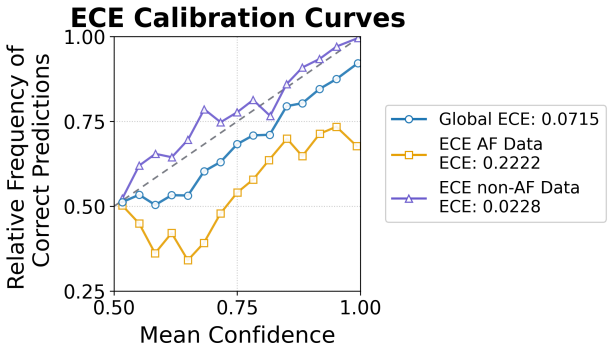

Fig. 2. Reliability diagrams for binary classifier trained to predict atrial fibrillation from raw PPG time series. Uncertainties are estimated with a range of techniques with Monte Carlo Dropout (MCD). The dashed line indicates perfect calibration. While the non-adaptive plot (blue) might indicate some degree of reliability, the adaptive plot shows significant discrepancy in the per class reliability. In particular, the minority class exhibits poor calibration. Figure created from data acquired from [12].

With that said, uncertainty reliability may not generalise well to all data classes without additional constraints. Even in the simple case of MCD applied to a binary classification task, uncertainty reliability can favour the dominant class as seen in [12], [22] and Fig. 2, which shows the global and adaptive reliability diagrams for a model trained to predict atrial fibrillation from raw PPG time series. While adaptivity can be assessed by plotting more standard reliability measures individually for each data class, issues with binning remain [12], and complicate analysis in cases where certain classes are highly underrepresented. While there are several variants of conformal prediction that aim to achieve adaptivity (per-class or otherwise) [35], [36], these are not well suited for data scarce applications common in medical settings. Therefore, there remains a strong desire for adaptive UQ methods that work in data scarce settings.

2) OOD data: Several works have shown that UQ techniques in deep learning do not generalise well to OOD data [32], [37]. This raises fundamental concerns with using estimated uncertainties to aid in decision making, as uncertainty estimates are often useful for highlighting examples that are likely to result in poor performance. In turn, these often correspond with those that are OOD relative to the training set. This motivates a need for thorough evaluations of the generalisability of uncertainty reliability. Yet, aside from some simple methods [38], there is a lack of thorough evaluation frameworks for assessing this. Greater effort should be placed towards understanding why there is some variability in the generalisability of different UQ frameworks used in deep learning, in characterising relevant sources of dataset shift, and developing methods to evaluate uncertainty generalisability. Previous efforts for benchmarking uncertainty reliability [39], [40] do not use adequate measures or factor in all relevant aspects related to the use case of the model.

D. General practical considerations

There are quite general challenges with uncertainty quantification in deep learning that also affect its utility in medical and other applications.

1) Model parameterisation: Many UQ techniques in deep learning are parameterisable, where uncertainty reliability is dependent on the chosen parameterisation (e.g. see examples from [41], and Fig. 3 that shows the reliability diagrams of classifiers trained to predict atrial fibrillation from raw PPG time series with MCD where each employs a different dropout rate). Yet, there is a lack of an efficient and generalisable methodology for choosing parameters that optimise reliability. While MCD can be adapted to optimise the dropout rate as part of the training objective [42], in general, a grid search over various parameterisations to optimise uncertainty reliability is recommended [16]. However, this can be expensive given model training times, and the number of parameters that need to be optimised [42]. This strategy is further complicated by the fact that some models may be used in a way such that they do not have a well defined quantitative notion of prediction error which complicates the formulation of measures for uncertainty reliability. As will be discussed, this is relevant in generative modelling, where uncertainty in an output may correlate with whatever notion of error is used, but may not factor in other contextual factors about how a model used for a downstream task may interpret image features that are actually relevant to the decision making process. In general, the lack of measures for individual reliability prevents the effective use of grid searches; though perhaps proxy measures can be used instead. The consequences of certain architectural choices are not well understood with respect to the optimal choice of hyperparameters, e.g. the placement of dropout for MCD [41] or the number of ensembles used in deep ensembles [43]; greater effort should be placed understanding how uncertainties are affected by these design choices.

2) Uncertainty disentanglement: It has been shown that common uncertainty disentanglement strategies in deep learning used for decomposing total uncertainty estimates into their constituent aleatoric and epistemic contributions are not

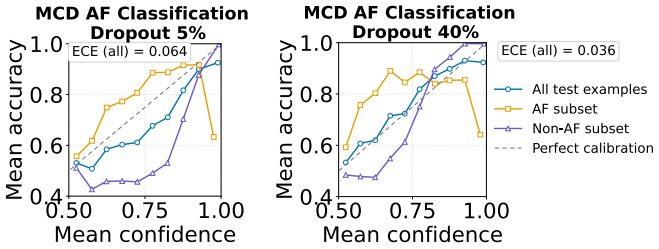

Fig. 3. Plot of reliability diagrams for binary classifier trained on raw PPG time series to predict AF classification. Uncertainties are estimated with Monte Carlo Dropout. Here, the chosen dropout rate has a considerable effect on the reliability of the uncertainties. The figure was created from data reported in [41].

effective, as evidenced by the significant correlation between the two sources [41], [44]. While in principle estimates of epistemic uncertainty may help inform design choices for model architecture, or aid in the curation of suitable training sets, in practice, disentangled uncertainty estimates can not be used for this purpose limiting the utility of the predicted uncertainties. This also limits the utility of disentangled estimates of aleatoric uncertainty for understanding the extent that the ill-posedness of the model's task or noise can have on the trustworthiness of predictions.

E. Generative modelling

Given restrictions placed on the use of real medical data and the labour involved with acquiring and collating data, there is considerable interest in using generative deep learning to procure more easily shareable/usable synthetic datasets. With that said, it is well known that generative deep learning models such as generative adversarial networks and diffusion models are prone to producing data with hallucinations and artefacts that may not occur in real data [45]. This can make them unsuitable for use as training data for downstream tasks or other purposes related to model development (e.g. synthetic benchmark datasets for evaluating model performance). Uncertainty quantification has been proposed as a means to filter low-quality generations [46], but assessing the reliability of estimates is complicated by the lack of a ground truth for each generation in many cases. In principle, it may be possible to conduct a coarse and heuristic analysis of uncertainty reliability using proxy measures for generated data quality (e.g. observe whether uncertainty correlates with image quality [46], [47]). However, image quality measures commonly used in generative modelling are not sufficient in this context as they do not factor in task-specific quality indicators relevant to the downstream task they are used for. E.g. when used as training data for deep learning models trained for ultimate use on real data, synthetic image quality is model architecture-dependent, and thus affects the downstream generalisability/reliability of models trained on this data. While in principle, the performance of a model trained on the downstream tasks using synthetic data may provide a more effective notion of image quality, this is an often expensive and impractical choice in many scenarios. Emerging decision-theoretic approaches aligned with this concept [48] may be suitable in cases where prediction error on the model outputs is ill-defined, but the consequences of using these outputs for decisions can be evaluated through a well-defined loss function.

In addition to the lack of appropriate context of how the generated data is used in practice, commonly used synthetic data quality metrics may also have their own intrinsic uncertainty. E.g. some quality metrics leverage auxiliary deep learning models to detect characteristic feature representations, and compare them with those detected from a reference set of real images (e.g. the Fréchet Inception Distance [49]). However, their ability to reliably detect these feature representations is inherently uncertain (depends on alignment between the data domains of the embedding model's training set, and the samples fed into it among other factors) that will influence the effectiveness of the metric; the consequences of this are under explored in the context of uncertainty evaluation. More clear and reliable notions of error in the generated images is needed to enable the effective use of uncertainty quantification in generative modelling.

F. Decision making

Commonly used uncertainty reliability measures in deep learning indicate whether the magnitude of uncertainties reflect the accuracy in the predicted measurand (quantity being measured). They do not consider the effect an uncertainty estimate may have on patient outcomes given how it is used within a given clinical decision making process/work flow. Therefore, generic frameworks for evaluating uncertainty reliability do not necessarily indicate the practical utility of uncertainty estimates when used in clinical work flows [24], [25]. Emerging decision theoretic frameworks for uncertainty quantification [25], [48] (discussed in Section IV-E) may improve the practical utility of estimates by considering the consequences of using them in some clinical decision making process. While this also applies to uncertainties from more generic measurement devices, this is particularly relevant to deep learning models given their ability to perform complex tasks.

V. Conclusions and Research Priorities

UQ techniques in deep learning aim to express the doubt in a given model output/prediction. However, there are inadequacies with common UQ techniques and reliability measures in deep learning that make it challenging to achieve and assess the extent of uncertainty reliability desired in medical contexts. To summarise, the main challenges are:

1) a lack of an effective means to incorporate relevant patient-specific contextual information for estimating/expressing uncertainty, or accounting for this in model optimisation schemes needed for complex/autonomous decision making,
2) a lack of UQ techniques and evaluation measures that effectively assess individual calibration and adaptivity in data scarce scenarios,

3) poor methodological frameworks for optimising training parameterisation and the evaluation of the generalisability of uncertainty reliability,
4) a lack of effective accuracy measures for evaluating generated data quality in medicine, and therefore uncertainty reliability for generative deep learning,
5) common uncertainty reliability measures do not consider the effect uncertainty predictions may have on patient outcomes given how they are used with respect to some clinical work flow.

For 1), as suggested in [7], deliberation with medical professionals is needed to establish contextual factors that may strongly influence patient outcomes for complex decision making [26]. Further effort should also be placed towards finding ways to quantify and incorporate this information into the model optimisation procedure, and ensuring predictions are output in a manner that enables them to be used appropriately in light of remaining limitations in incorporating these factors, e.g. low uncertainty estimates could be expressed as a recommendation for further investigation of possible contextual information to prevent poor outcomes of planning treatments based on the model output alone. Establishing these factors and useful expressions of uncertainty will be task and model specific.

For 2), greater emphasis should be placed on developing and testing UQ techniques that directly optimise for adaptivity, and individual reliability while also catering to the practical realities of medical data (significant class imbalance, and smaller dataset sizes). Conformal prediction may offer promising paths towards achieving this [28], while intrinsic modelling approaches seem less flexible in this respect. Some evidential deep learning approaches have been developed to address class imbalance, but these should be tested at scale on realistic problems [31]; this also applies to other variants of popular UQ techniques that aim to address individual reliability [30] or other weaknesses. Novel quantitative measures for individual reliability are also needed, along with some notion of uncertainty in the measures (e.g. introduced by binning limitations or otherwise). While some notions of individual calibration can be assessed with an average difference between prediction error and uncertainty, this fails to capture sharpness which is a critical aspect of uncertainty reliability.

For 3), greater effort should be placed towards developing general frameworks for assessing the generalisability of uncertainty reliability, including sensible criteria/measures for defining the extent to which test data is OOD relative to training data. Here it is important to stratify data based on factors relevant to the use case (e.g. generalisation for minority data classes vs. majority classes). Additional efforts should be placed towards improving methods for optimising model parameterisations, which should involve a better understanding of how these parameters affect uncertainty reliability. Greater understanding of the various dependencies can help streamline hyperparameter choice. While some efforts to benchmark uncertainty reliability have been conducted [39], [40], the use of common evaluation measures, and lack of stratification to assess reliability with respect to the use case of the model falls short of providing a generalisable framework relevant to medical applications.

For 4), there is a lack of rigorously formulated procedures for evaluating uncertainty reliability in cases where there are no ground truths, and therefore, no clear notion of prediction error. While proxy-measures for generated data quality could be used to facilitate a coarse reliability analysis [47], more work should be placed towards establishing their trustworthiness. Furthermore, quality measures should factor in the context of how the data is used in whatever downstream task they are generated for.

For 5), ultimately, uncertainty estimates are useful if they indicate whether the use of the prediction in some decision making process/work flow will result in the desired outcome; emerging decision-theoretic frameworks catering to this view [48] may improve the practical utility of estimates. They may also more readily apply to tasks that lack a well defined notion of prediction error for the model's outputs (e.g. generative modelling - point 4) but may instead have a well defined loss on the consequences from the decision made using the output.

Progress in these areas will help strengthen the effective use of uncertainties in medicine, and ultimately help realise the potential deep learning could have for improving patient outcomes.

REFERENCES

[1] Thuy D Nguyen, Christopher M Whaley, Kosali Simon, Neil Mehta, Hao Yu, Ryan K McBain, Ateev Mehrotra, and Jonathan H Cantor. Adoption of artificial intelligence in the health care sector. In *JAMA Health Forum*, volume 6, pages e255029–e255029. American Medical Association, 2025.

[2] Sidra Naveed, Gunnar Stevens, and Dean Robin-Kern. An overview of the empirical evaluation of explainable AI (XAI): A comprehensive guideline for user-centered evaluation in XAI. *Applied Sciences*, 14(23):11288, 2024.

[3] Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 2022.

[4] Stefan Haufe, Rick Wilming, Benedict Clark, Rustam Zhumagambetov, Danny Panknin, and Ahcène Boubekki. Explainable AI needs formal notions of explanation correctness. *arXiv preprint arXiv:2409.14590*, 2024.

[5] Jean-Christophe Bélisle-Pipon. Commentary: Implications of causality in artificial intelligence. *Frontiers in Artificial Intelligence*, 7:1488359, 2025.

[6] Agustina D Saenz, Zach Harned, Oishi Banerjee, Michael D Abràmoff, and Pranav Rajpurkar. Autonomous AI systems in the face of liability, regulations and costs. *NPJ digital medicine*, 6(1):185, 2023.

[7] Sylvie Delacroix, Diana Robinson, Umang Bhatt, Jacopo Domenicucci, Jessica Montgomery, Gaël Varoquaux, Carl Henrik Ek, Vincent Fortuin, Yulan He, Tom Diethe, et al. Beyond quantification: Navigating uncertainty in professional AI systems. *RSS: Data Science and Artificial Intelligence*, 1(1):udaf002, 2025.

[8] Kyle Karches. Hermeneutics as impediment to AI in medicine. *Theoretical Medicine and Bioethics*, pages 1–19, 2025.

[9] Pascal Pernot. Calibration in machine learning uncertainty quantification: beyond consistency to target adaptivity. *APL Machine Learning*, 1(4), 2023.

[10] I ISO. and BIPM OIML. *Guide to the Expression of Uncertainty in Measurement*. Aenor Madrid, Spain, 1993.

[11] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[12] Ciaran Bench, Oskar Pfeffer, Vivek Desai, Mohammad Moulaeifard, Loïc Coquelin, Peter H Charlton, Nils Strodthoff, Nando Hegemann, Philip J Aston, and Andrew Thompson. A systematic evaluation of uncertainty quantification techniques in deep learning: a case study in photoplethysmography signal analysis. *arXiv preprint arXiv:2511.00301*, 2025.

[13] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[14] Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222*, 2023.

[15] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on Bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

[16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[17] Shivam Pal, Aishwarya Gupta, Saqib Sarwar, and Piyush Rai. Simple and scalable federated learning with uncertainty via Improved Variational Online Newton. In *OPT 2024: Optimization for Machine Learning*, 2024.

[18] Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, 2024.

[19] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[20] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[21] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268, 2007.

[22] Ciaran Bench, Nils Strodthoff, Mohammad Moulaeifard, Philip James Aston, and Andrew Thompson. Towards trustworthy atrial fibrillation classification from wearables data: quantifying model uncertainty. *Computing in cardiology*, 51, 2024.

[23] Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 22(15):5540, 2022.

[24] Richard D Riley, Gary S Collins, Laura Kirton, Kym IE Snell, Joie Ensor, Rebecca Whittle, Paula Dhiman, Maarten Van Smeden, Xiaoxuan Liu, Joseph Alderman, et al. Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches. *BMJ*, 388, 2025.

[25] Shayan Kiyani, George Pappas, Aaron Roth, and Hamed Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. *arXiv preprint arXiv:2502.02561*, 2025.

[26] M Alsuleman, P Duncan, and A Thompson. Screening of atrial fibrillation using wearable PPG devices–a trustworthy and safe AI life cycle case study. 2025.

[27] Mark Levene, Tameem Adel, Moulham Alsuleman, Indhu George, Padmini Krishnadas, Keith Lines, Yuhui Luo, Ian Smith, and Paul Duncan. A life cycle for trustworthy and safe artificial intelligence systems. 2024.

[28] Tapabrata Chakraborti, Christopher RS Banerji, Ariane Marandon, Vicky Hellon, Robin Mitra, Brieuc Lehmann, Leandra Bräuninger, Sarah McGough, Cagatay Turkay, Alejandro F Frangi, et al. Personalized uncertainty quantification in artificial intelligence. *Nature Machine Intelligence*, 7(4):522–530, 2025.

[29] Shengjia Zhao, Tengyu Ma, and Stefano Ermon. Individual calibration with randomized forecasting. In *International Conference on Machine Learning*, pages 11387–11397. PMLR, 2020.

[30] Hassan Gharoun, Mohammad Sadegh Khorshidi, Kasra Ranjbarigderi, Fang Chen, and Amir H Gandomi. Uncertainty-aware post-hoc calibration: Mitigating confidently incorrect predictions beyond calibration metrics. *arXiv preprint arXiv:2510.17915*, 2025.

[31] Tong Xia, Ting Dang, Jing Han, Lorena Qendro, and Cecilia Mascolo. Uncertainty-aware health diagnostics via class-balanced evidential deep learning. *IEEE Journal of Biomedical and Health Informatics*, 28(11):6417–6428, 2024.

[32] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[33] Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. Uncertainty calibration error: A new metric for multi-class classification, 2021. OpenReview; ICLR 2021 submission. Version posted: 2023-05-05.

[34] Edward S Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology (1962-1982)*, 8(6):985–987, 1969.

[35] Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

[36] Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126, 03 2025.

[37] Christian Tomani, Sebastian Gruber, Muhammed Ebrar Erdem, Daniel Cremers, and Florian Buettner. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10132, 2021.

[38] A Michael Carrell, Neil Mallinar, James Lucas, and Preetum Nakkiran. The calibration generalization gap. *arXiv preprint arXiv:2210.01964*, 2022.

[39] Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael W Dusenberry, Sebastian Farquhar, Qixuan Feng, Angelos Filos, Marton Havasi, Rodolphe Jenatton, et al. Uncertainty baselines: Benchmarks for uncertainty & robustness in deep learning. *arXiv preprint arXiv:2106.04015*, 2021.

[40] Michael Kirchhof, Bálint Mucsányi, Seong Joon Oh, and Enkelejda Kasneci. Url: A representation learning benchmark for transferable uncertainty estimates. *Advances in Neural Information Processing Systems*, 36:13956–13980, 2023.

[41] Ciaran Bench, Vivek Desai, Mohammad Moulaeifard, Nils Strodthoff, Philip Aston, and Andrew Thompson. Uncertainty quantification with approximate variational learning for wearable photoplethysmography prediction tasks. *arXiv preprint arXiv:2505.11412*, 2025.

[42] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *Advances in neural information processing systems*, 30, 2017.

[43] Ekaterina Lobacheva, Nadezhda Chirkova, Maxim Kodryan, and Dmitry P Vetrov. On power laws in deep ensembles. *Advances In Neural Information Processing Systems*, 33:2375–2385, 2020.

[44] Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Advances in neural information processing systems*, 37:50972–51038, 2024.

[45] Matthew Tivnan, Siyeop Yoon, Zhennong Chen, Xiang Li, Dufan Wu, and Quanzheng Li. Hallucination index: An image quality metric for generative reconstruction models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 449–458. Springer, 2024.

[46] Metod Jazbec, Eliot Wong-Toi, Guoxuan Xia, Dan Zhang, Eric Nalisnick, and Stephan Mandt. Generative uncertainty in diffusion models. *arXiv preprint arXiv:2502.20946*, 2025.

[47] Ciaran Bench, Emir Ahmed, and Spencer A Thomas. Trustworthy image-to-image translation: evaluating uncertainty calibration in unpaired training scenarios. *arXiv preprint arXiv:2501.17570*, 2025.

[48] Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark van der Wilk, Adam Foster, and Tom Rainforth. Rethinking aleatoric and epistemic uncertainty. *arXiv preprint arXiv:2412.20892*, 2024.

[49] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in neural information processing systems*, 30, 2017.