Analyzing The Role Of Model Uncertainty For Electronic Health Records

Michael W. Dusenberry*†
Google Brain

Dustin Tran
Google Brain

Edward Choi
Google Health

Jonas Kemp
Google Health

Jeremy Nixon
Google Brain

Ghassen Jerfel
Google Brain, Duke University

Katherine Heller
Google Brain

Andrew M. Dai
Google Health

ABSTRACT

In medicine, both ethical and monetary costs of incorrect predictions can be significant, and the complexity of the problems often necessitates increasingly complex models. Recent work has shown that changing just the random seed is enough for otherwise well-tuned deep neural networks to vary in their individual predicted probabilities. In light of this, we investigate the role of model uncertainty methods in the medical domain. Using recurrent neural network (RNN) ensembles and various Bayesian RNNs, we show that population-level metrics, such as AUC-PR, AUC-ROC, log-likelihood, and calibration error, do not capture model uncertainty. Meanwhile, the presence of significant variability in patient-specific predictions and optimal decisions motivates the need for capturing model uncertainty. Understanding the uncertainty for individual patients is an area with clear clinical impact, such as determining when a model decision is likely to be brittle. We further show that RNNs with only Bayesian embeddings can be a more efficient way to capture model uncertainty compared to ensembles, and we analyze how model uncertainty is impacted across individual input features and patient subgroups.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; **Neural networks**; **Uncertainty quantification**; • **Applied computing → Life and medical sciences**.

KEYWORDS

uncertainty, neural networks, Bayesian deep learning, electronic health records

---

*Work completed as a Google AI Resident.

†Correspondence to: dusenberrymw@google.com.

1 INTRODUCTION

Machine learning has found great and increasing levels of success in the last several years on many well-known benchmark datasets. This has led to a mounting interest in non-traditional problems and domains, each of which bring their own requirements. In medicine specifically, individualized predictions are of great importance to the field [5], and there can be severe costs for incorrect decisions due to the risk to human life and associated ethical concerns [10].

Existing state-of-the-art approaches using deep neural networks in medicine often make use of either a single model or an average over a small ensemble of models, focusing on improving the accuracy of probabilistic predictions [4, 14, 26, 35]. These works, while focusing on capturing the data uncertainty, do not address the *model* uncertainty that is inherent in fitting deep neural networks [16, 22]. For example, when predicting patient mortality in an ICU setting, existing approaches might be able to achieve high AUC-ROC, but will be unable to differentiate between patients for whom the model is *certain* about its probabilistic prediction, and those for whom the model is fairly *uncertain*.

In this paper, we examine the use of model uncertainty specifically in the context of predictive medicine. Model uncertainty has made many methodological advances in recent years—including reparameterization-based variational Bayesian neural networks [3, 7, 19, 21], Monte Carlo dropout [8], deep ensembles and efficient alternatives [20, 33], and function priors [9, 13, 22]. Deep neural networks combined with advanced model uncertainty methods can directly impact clinical care by answering several questions that naturally occur in predictive medicine:

- How do the realized functions in any of the approaches, such as individual models in the ensemble approach, compare in terms of population-level metric performance such as AUC-PR, AUC-ROC, or log-likelihood?
- If and how does model uncertainty assist in calibrating predictions?
- How does model uncertainty change across different patient subgroups, in terms of ethnicity, gender, age, or length of stay?

- How do various feature values contribute towards model uncertainty?
- How does model uncertainty affect optimal decisions made under a given clinically-relevant cost function?

Contributions. Using sequence models on the MIMIC-III [15] and eICU [25] clinical datasets, we make several important findings. For the ensembling approach of quantifying model uncertainty, we find that the models within the ensemble can collectively exhibit a wide variability in predicted probabilities for some patients, despite being well-calibrated and having *nearly identical dataset-level metric performance*. We find that this even extends into the space of optimal decisions. That is, models with nearly equivalent metric performance can disagree significantly on the final decision, thus transforming an "optimal" decision into a random variable. Significant variability in patient-specific predictions and decisions can be an indicator of when a model decision is likely to be brittle, and we show that using a single model or an average over models can mask this information. This motivates the importance of model uncertainty for clinical decision systems. Given this, we proceed with an analysis over different clinical tasks and datasets, looking at how model uncertainty is impacted across individual input features and patient subgroups. We then show that models with Bayesian embeddings can be a more efficient way to capture model uncertainty compared to deep ensembles.

2 BACKGROUND

2.1 Data Uncertainty

Data uncertainty can be viewed as uncertainty regarding a given outcome due to incomplete information, and is also known as "output uncertainty", "noise", or "risk" [18]. This uncertainty is represented by the predictive distribution

$$y \sim p(y|\mathbf{x}) \tag{1}$$

for the outcome y given inputs $\mathbf{x}$. In a learning scenario, we could define a function $f(\mathbf{x}, \mathbf{w})$ with learnable parameters $\mathbf{w}$ that outputs a parameterization of the predictive distribution $p(y|\mathbf{x}, \mathbf{w})$, which is now conditioned on $\mathbf{w}$. For binary tasks, the predictive distribution equates to a Bernoulli distribution, which is parameterized by a single probability value. More specifically, for the binary case, this can be described as

$$\begin{aligned} \lambda &= f(\mathbf{x}, \mathbf{w}) \\ y &\sim \text{Bernoulli}(\lambda), \end{aligned} \tag{2}$$

where the model f, as a function of the inputs $\mathbf{x}$ and parameters $\mathbf{w}$, outputs the parameter λ (a vector of length one) for the Bernoulli distribution representing the conditional distribution $p(y|\mathbf{x}, \mathbf{w})$ for the outcome y. For multiclass tasks, the predictive distribution takes the form of a Multinomial distribution with a single trial (and parameterized by a vector λ), and for regression tasks, one could use a continuous distribution such as a Gaussian.

2.2 Model Uncertainty

Model uncertainty can be viewed as uncertainty regarding the true function underlying the observed process [2]. For a learned function $f(\mathbf{x}, \mathbf{w})$ of inputs $\mathbf{x}$ and parameters $\mathbf{w}$, this uncertainty is

represented by a distribution over functions [2, 34]

$$f \sim \mathcal{G}(f) \tag{3}$$

which is often induced by a distribution over the function parameters [2, 3, 7, 34]

$$\mathbf{w} \sim p(\mathbf{w}). \tag{4}$$

Because different functions can yield different predictive distributions, a distribution over functions leads to a distribution over predictive distributions, representing *disagreement* due to model uncertainty. We can see this more formally by defining a function $g_{\mathbf{x}}(\mathbf{w}) = f(\mathbf{x}, \mathbf{w})$ for a given input $\mathbf{x}$, and then viewing this as a change of variables from $\mathbf{w}$ to λ,

$$\begin{aligned} \lambda &= g_{\mathbf{x}}(\mathbf{w}) = f(\mathbf{x}, \mathbf{w}) \\ \mathcal{P}(\lambda \in \mathcal{A} \mid \mathbf{x}) &= \int_{\{\mathbf{w}|g_{\mathbf{x}}(\mathbf{w})=\lambda \in \mathcal{A}\}} p(\mathbf{w}) \, \mathrm{d}\mathbf{w} \\ p(\lambda|\mathbf{x}) &= \frac{\mathrm{d}}{\mathrm{d}\lambda} \int_{\{\mathbf{w}|g_{\mathbf{x}}(\mathbf{w})\le\lambda\}} p(\mathbf{w}) \, \mathrm{d}\mathbf{w}, \end{aligned} \tag{5}$$

where the distribution over $\mathbf{w}$ is transformed into a distribution over λ (conditioned on $\mathbf{x}$). Thus, there is an induced distribution over the parameters of the predictive distribution due to uncertainty in the function space. For binary tasks, this would equate to a distribution of plausible probability values for a Bernoulli distribution.

We can then write down the final, marginalized predictive distribution

$$\begin{aligned} p(y|\mathbf{x}) &= \int p(y|\mathbf{x}, \mathbf{w})p(\mathbf{w}) \, \mathrm{d}\mathbf{w} \\ &= \int p(y|\mathbf{x}, \lambda)p(\lambda|\mathbf{x}) \, \mathrm{d}\lambda \end{aligned} \tag{6}$$

in two equivalent forms. Importantly, by considering the distribution $p(\lambda|\mathbf{x})$ before marginalizing, we can compute two quantities of interest: the expected value $\mathbb{E}_{\lambda \sim p(\lambda|\mathbf{x})}[\lambda|\mathbf{x}]$ (which is used in the marginalization of equation 6), and a measure of *disagreement* (or uncertainty due to model uncertainty) such as the variance $\text{Var}[\lambda|\mathbf{x}]$. It is also important to note that the variance in the final, marginalized *predictive distribution* $p(y|\mathbf{x})$ will include *both* data uncertainty *and* model uncertainty sources, but it is not possible to distinguish the two from that marginalized distribution alone (and thus $p(\lambda|\mathbf{x})$ is needed).

For the remainder of the paper, we will use the phrase *predictive uncertainty distribution* to refer to the distribution $p(\lambda|\mathbf{x})$ over the parameter(s) of the predictive distribution as induced by the uncertainty over model parameters.

2.3 Calibration

A model is said to be perfectly calibrated if, for all examples for which the model produces the same prediction p for some outcome, the percentage of those examples truly associated with the outcome is equal to p, across all values of p. If a model is systematically over- or under-confident, it can be difficult to reliably use its predicted probabilities for decision making. The expected calibration error (ECE) metric [23] is one tractable way to approximate the calibration of a model given a finite dataset. ECE computes a weighted
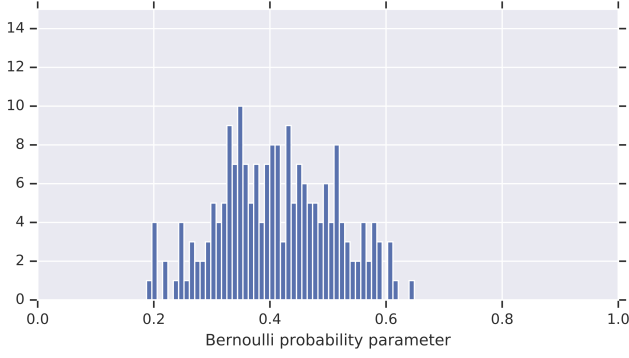
Figure 1: A histogram of predictions from M deterministic recurrent neural network (RNN) models trained with different random seeds for a single intensive care unit (ICU) patient's probability of mortality. As shown here, model uncertainty can cause high disagreement between individual models in an ensemble regarding the correct predictive distribution for a given patient. This is not captured when using a single model or an average over an ensemble.
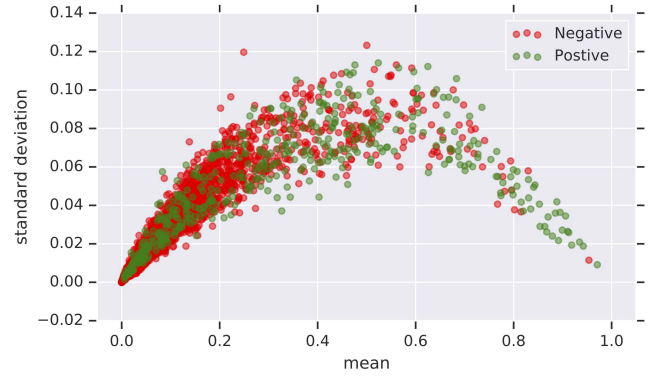

Figure 2: A plot of the mean versus standard deviation of the predictive uncertainty distributions of the deterministic ensemble for positive and negative patients in the validation set. We find that the standard deviations do not form a simple linear relationship with the mean. For reference, we note that the variances of the distributions are generally lower than that of a Bernoulli distribution's variance curve.

average of the calibration error across bins, and is defined as

$$\text{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} \left| \text{acc}(b) - \text{conf}(b) \right|,$$

where n_b is the number of predictions in bin b, N is the total number of data points, and $\text{acc}(b)$ and $\text{conf}(b)$ are the accuracy and confidence of bin b, respectively. Recent work [12] has shown that modern deep neural networks (NNs) tend to be poorly calibrated.

2.4 Deep Ensembles

Deep ensembles [20] is a method for quantifying model uncertainty. In this approach, an ensemble of M deterministic[1] NNs is trained by varying only the random seed of an otherwise well-tuned set of hyperparameters. Given this ensemble, a prediction $\boldsymbol{\lambda}^{(m)}$ can be made with each model m for a given input $\mathbf{x}$, where (for a binary task) each prediction is the probability parameter for the Bernoulli distribution over the outcome. The set of M probabilistic predictions $\{\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}, \dots, \boldsymbol{\lambda}^{(M)}\}$ for the same example can then be viewed as samples from the distribution $p(\boldsymbol{\lambda}|\mathbf{x})$ (equation 5), where this distribution represents disagreement, or uncertainty due to model uncertainty. In this work, we make use of deep ensembles of RNNs to model sequential patient data.

2.5 Bayesian RNNs

Bayesian RNNs [7] are RNNs with a prior distribution $p(\mathbf{w})$ placed over the parameters $\mathbf{w}$ of the model. This allows us to express model uncertainty as uncertainty over the true values for the parameters in the model, *i.e.*, "weight uncertainty" [3]. By introducing a distribution over all, or a subset, of the weights in the model, we can induce different functions, and thus different outcomes, through

realizations of different weight values via draws from the posterior distributions over those weights. This allows us to empirically capture model uncertainty in the predictive uncertainty distribution $p(\boldsymbol{\lambda}|\mathbf{x})$ by drawing M samples from a single Bayesian RNN for a given example. In this work, we make use of various Bayesian RNN variants by placing priors on different subsets of the parameters.

3 MEDICAL UNCERTAINTY

3.1 Clinical Tasks

We demonstrate results on both binary and multiclass clinical tasks using multiple electronic health record (EHR) datasets. In terms of data, we use

(1) Medical Information Mart for Intensive Care (MIMIC-III) [15], and
(2) eICU Collaborative Research Database (eICU) [25],

both of which are publicly available EHR datasets. MIMIC-III is collected from 46,520 patients admitted to ICUs at Beth Israel Deaconess Medical Center, where 9,974 expired during the encounter (*i.e.*, 1:4 ratio between positive and negative samples). The eICU dataset is collected from over 200,000 admissions to ICUs across the United States. In terms of tasks, for MIMIC-III we study

(1) binary in-patient mortality prediction, and
(2) multiclass diagnosis prediction at discharge.

For the multiclass diagnosis prediction, we use the single-level Clinical Classifications Software (CCS) code system. For the eICU dataset, we study the binary in-patient mortality prediction task as well, allowing us to demonstrate that our findings generalize to additional datasets.

---

[1]We use the term "deterministic" to refer to the usual setup in which we optimize the parameter values of our function directly, yielding a trained model with fixed parameter values at test time.

3.2 Models

Similar to Rajkomar et al. [26], we train deep RNNs for our clinical tasks. Each of our models embeds and aggregates a patient's sequential features (*e.g.* medications, lab measures, clinical notes) and global contextual features (*e.g.* gender, age), feeds them to one or more long short-term memory (LSTM) layers [28], and follows that with hidden and output affine layers. More specifically, sequential embeddings are bagged into 1-day blocks, and fed into one or more LSTM layers. The final time-step output of the LSTM layers is concatenated with the contextual embeddings and fed into a hidden dense layer, and the output of that layer is then fed into an output dense layer yielding the parameterization $\boldsymbol{\lambda}$ for a predictive distribution $p(y|\mathbf{x}, \mathbf{w})$. A ReLU non-linearity is used between the hidden and output dense layers, and default initializers in tf.keras.layers.* are used for all deterministic layers. More details on the training setup can be found in the Appendix and in the code[2]

Existing deep learning approaches in predictive medicine focus on capturing data uncertainty, namely accurately predicting the predictive distribution $p(y|\mathbf{x})$ of a patient outcome (*i.e.*, how likely is the patient to expire?). This work, on the other hand, also focuses on addressing the model uncertainty aspect of deep learning, namely the distribution over equally-likely predictive distributions (*i.e.*, are there alternative predictive distributions, and if so, how diverse are the distributions?).

3.3 Choice of Uncertainty Methods

To quantify model uncertainty for clinical tasks, we explore the use of deep RNN ensembles and various Bayesian RNNs. For the deep ensembles approach, we optimize for the ideal hyperparameter values for our RNN model via black-box Bayesian optimization [11], and then train M replicas of the best model. Only the random seed differs between the replicas. At prediction time, we make predictions with each of the M models for each patient. The full list of hyperparameters and the specific hyperparameter values for all models can be found in the Appendix.

For the Bayesian RNNs, we train a single model, and then draw M samples from it at prediction time. To train the Bayesian RNN, we take a variational inference approach by adapting our RNN to use factorized weight posteriors

$$q(\mathbf{w}|\boldsymbol{\theta}) = \prod_i q(\mathbf{w}^{(i)}|\boldsymbol{\theta}^{(i)}),$$

where each weight tensor $\mathbf{w}^{(i)}$ in the model is represented by a normal distribution with learnable mean and diagonal covariance parameters represented collectively as $\boldsymbol{\theta}^{(i)}$. Normal distributions with zero mean and tunable standard deviation are used as weight priors $p(\mathbf{w}^{(i)})$. We train our models by minimizing the Kullback-Leibler (KL) divergence

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \mathrm{KL}[q(\mathbf{w}|\boldsymbol{\theta}) \,\|\, p(\mathbf{w}|\mathbf{y}, \mathbf{X})] \\ &\propto \mathrm{KL}[q(\mathbf{w}|\boldsymbol{\theta}) \,\|\, p(\mathbf{w})] - \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\theta})}[\ln p(\mathbf{y}|\mathbf{X}, \mathbf{w})]\end{aligned} \quad (7)$$

between the approximate weight posterior $q(\mathbf{w}|\boldsymbol{\theta})$ and the true, but unknown posterior $p(\mathbf{w}|\mathbf{y}, \mathbf{X})$. Overall, this equates to minimizing an expectation over the usual negative log likelihood term, $\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\theta})}[\ln p(\mathbf{y}|\mathbf{X}, \mathbf{w})]$, plus a KL divergence regularization term. To

---

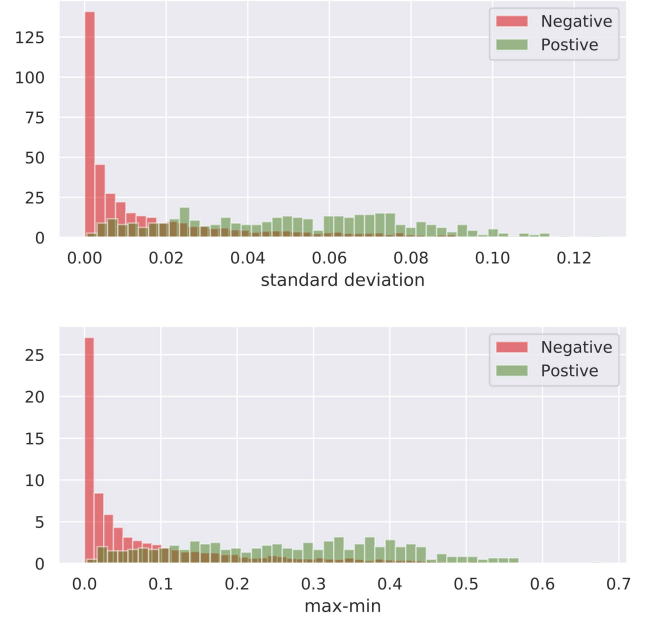[2]Code can be found at https://github.com/Google-Health/records-research.

Figure 3: Top: A histogram of the standard deviations of the $p(\boldsymbol{\lambda}|\mathbf{x})$ distributions for all patients in the test set. Bottom: The same setup, but looking at differences between the maximum and minimum values of those $p(\boldsymbol{\lambda}|\mathbf{x})$ distributions. Together, this shows that there is wide variability in predicted probabilities for some patients, and that negative patients have less variability on average.

easily shift between the deterministic RNN and various Bayesian RNN models, we make use of the Bayesian Layers [32] abstractions.

3.4 Optimal Decisions via Sensitivity Requirements

The key desire in clinical practice is to make a decision based on the model's predicted probability and its associated uncertainty. Given a set of potential outcomes $y_k \in \{1, \dots, K\}$ for K classes, a set of conditional probabilities $p(y_k|\mathbf{x})$ for the given outcomes, and the associated costs L_{kj} for predicting class j when the true class is k, an optimal decision can be determined by minimizing the expected decision cost

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(y_k|\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \quad (8)$$

with respect to $\mathcal{R}_j$, where $\mathcal{R}_j$ is the decision region for assigning example $\mathbf{x}$ to class j, and $p(\mathbf{x})$ is the density of $\mathbf{x}$ [2].

Designing elaborate decision cost functions for clinical applications is an interesting but difficult task, as it requires expert knowledge of the prediction target, cost-benefit analysis, and medical resource allocation. Fortunately, we can use a clinically relevant alternative, which is the *sensitivity requirement*. Often in clinical research, certain sensitivity (*i.e.*, recall) levels are desirable when making predictions in order for a model to be clinically relevant [6, 27, 29–31]. The goal in such cases is to maximize the precision

while still maintaining the desired sensitivity level. Viewed as a decision cost function, the cost is infinite if the recall is below the target level, and is otherwise minimized as the precision is increased, where the optimized parameter is a global probability threshold $t^{(m)}$ for a given model m.

For each of the M models in our ensemble, we can optimize the sensitivity-based decision cost function and make optimal decisions for all examples. Thus, for each example, there will be a set of M optimal decisions, which can be represented as a distribution. That is, from this viewpoint, the optimal decision d for an example $\mathbf{x}$ can be represented as a random variable

$$d \sim p(d|\mathbf{x}), \tag{9}$$

which, for a binary task, can be approximated as

$$\phi = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}(\boldsymbol{\lambda}^{(m)} \geq t^{(m)}))$$
$$d \sim \text{Bernoulli}(\phi), \tag{10}$$

where $\boldsymbol{\lambda}^{(m)} \geq t^{(m)}$ is the decision function for model m, and ϕ is the percentage of model agreement.

This simply represents the propagation of uncertainty over functions into uncertainty over predictive distributions (equation 5), and, in turn, into uncertainty over optimal decisions. That is, different equally-likely functions could yield different values for $\boldsymbol{\lambda}$ and thus different predictive distributions $p(y|\mathbf{x}, \mathbf{w})$ for a given example, which could lead to different optimal decisions for that example. In the same way that we could represent a set of functions as a distribution over functions, we could represent a set of predictive distributions as a distribution over predictive distributions, and we could represent a set of optimal decisions as a distribution over optimal decisions. As stated previously, the variance of these distributions represents *disagreement*, *i.e.*, uncertainty due to model uncertainty.

4 EXPERIMENTS

We perform four sets of experiments. First, in order to demonstrate the importance of quantifying uncertainty in predictive medicine, we examine individual models in the RNN ensemble in terms of predictive metrics, calibration, uncertainty distributions, and decision-making. Second, we examine multiple variants of Bayesian RNNs to understand where uncertainty in the model matters most, comparing them with their deterministic ensemble counterpart. Third, we use the deterministic RNN ensemble to examine uncertainty across different patient subgroups. Finally, we analyze the Bayesian RNN with embedding distributions to examine uncertainty across individual features.

4.1 When Do We Observe Uncertainty?

Clinical Metrics. For our clinical tasks, we first measure the dataset-level metrics:

- area under the precision-recall curve (AUC-PR) (*binary tasks*),
- area under the receiver operating characteristic curve (AUC-ROC) (*binary tasks*),
- top-5 recall (*multiclass tasks*),
- top-5 precision (*multiclass tasks*),
- top-5 F1 (*multiclass tasks*),

Table 1: Dataset-level metrics for the MIMIC-III binary mortality and multiclass CCS prediction tasks across $M = 200$ models in the deterministic RNN ensemble. Metrics are computed for each model within the ensemble, and means and standard deviations across models are reported. Individual models are nearly identical in terms of dataset-level performance across both tasks, but selecting a single model would remove the model uncertainty information such as that visualized in Figure 1.

	METRIC	VALIDATION	TEST
MORTALITY	AUC-PR ↑	0.4496 (0.0025)	0.3886 (0.0059)
	AUC-ROC ↑	0.8753 (0.0019)	0.8623 (0.0031)
	NEG. LOG-LIKELIHOOD ↓	0.2037 (0.0030)	0.2088 (0.0038)
	ECE ↓	0.0176 (0.0040)	0.0162 (0.0043)
	ACE ↓	0.0210 (0.0042)	0.0233 (0.0057)
CCS DIAGNOSIS	TOP-5 RECALL ↑	0.7126 (0.0071)	0.7090 (0.0088)
	TOP-5 PRECISION ↑	0.1425 (0.0014)	0.1418 (0.0018)
	TOP-5 F1 ↑	0.2375 (0.0024)	0.2363 (0.0029)
	NEG. LOG-LIKELIHOOD ↓	2.2738 (0.0330)	2.3338 (0.0434)
	ECE ↓	0.0446 (0.0072)	0.0499 (0.0082)
	ACE ↓	4.219e−3 (7.31e−8)	4.219e−3 (7.61e−8)

- held-out negative log-likelihood (*all tasks*),
- ECE (*all tasks*) [23], and
- adaptive calibration error (ACE) (*all tasks*) [24].

Table 1 shows the performance on the MIMIC-III binary mortality and multiclass CCS multiclass tasks averaged over individual models in our deterministic RNN ensemble, with the standard deviation over models in the parentheses. Interestingly, individual models are overall well-calibrated and *nearly equivalent in terms of likelihood and metric performance*. If we were to choose only one model in practice based on the dataset-level metrics, it is highly likely any of the models in the ensemble could be selected. Importantly, if we only used a single model, we would lose the model uncertainty information (as noted in Section 2.2).

Predictive Uncertainty Distributions & Statistics. Knowing that the models in our ensemble are well-calibrated and effectively equivalent in terms of performance, we turn to making predictions for individual examples. Figure 1 visualizes the predictive uncertainty distribution for a single patient on the mortality task using the deterministic RNN ensemble. We find that there is a wide variability in predicted Bernoulli probabilities for some patients (with spreads as high as 57.5%). As noted in Section 2.2, this variability represents our uncertainty associated with determining the correct predictive distribution $p(y|\mathbf{x}, \mathbf{w})$ for the given patient. Marginalizing over this uncertainty with respect to $\mathbf{w}$ will yield the current best estimate for $p(y|\mathbf{x})$, but the estimate could be improved through the acquisition of more training examples similar to the current patient. Ignoring the variance $\text{Var}[\boldsymbol{\lambda}|\mathbf{x}]$ through the use of either a single model or an average over models without also conveying the original variance is likely detrimental since it is not possible to distinguish between data uncertainty and model uncertainty from that marginalized distribution $p(y|\mathbf{x})$ alone, and thus it prevents a physician from being
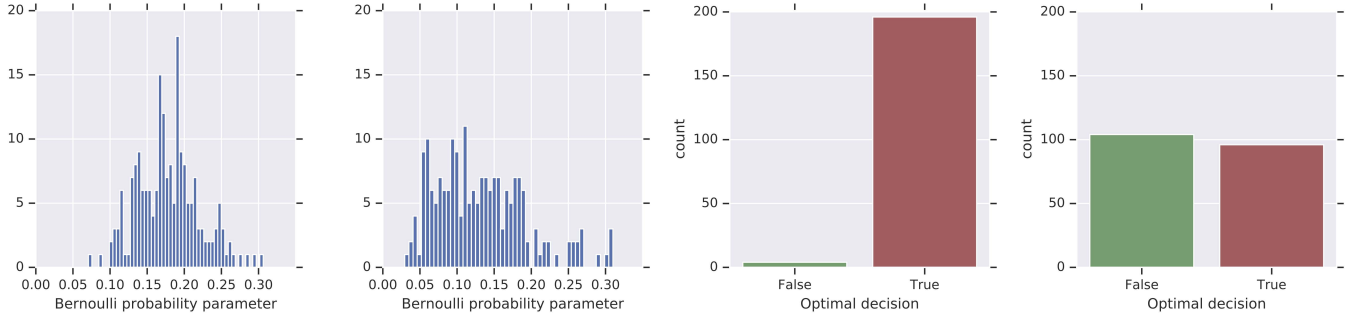
Figure 4: *Left Two:* **Histograms representing the different mortality predictive distributions produced by the deterministic ensemble for two patients in the validation set.** *Right Two:* **The corresponding optimal decision distributions, with "True" corresponding to a prediction of mortality and "False" corresponding to the opposite. For one patient, the ensemble members are in agreement about the optimal decision, while for the other patient there is high disagreement due to model uncertainty.**

Table 2: Metrics for marginalized predictions on the MIMIC-III and eICU mortality tasks given $M = 200$ models in the deterministic RNN ensemble, and $M = 200$ samples from each of the Bayesian RNN models. 95% confidence intervals are computed via validation and test set bootstrapping with 1000 bootstrap sets.

	Model	Val. AUC-PR ↑	Val. AUC-ROC ↑	Val. NLL ↓	Test AUC-PR ↑	Test AUC-ROC ↑	Test NLL ↓
MIMIC-III	Deterministic Ensemble	0.4564 (±1e−3)	0.8774 (±5e−4)	0.1999 (±5e−4)	0.3921 (±1e−3)	0.8643 (±5e−4)	0.2051 (±5e−4)
	Bayesian Embeddings	0.4580 (±1e−3)	0.8776 (±4e−4)	0.2002 (±4e−4)	0.3977 (±2e−3)	0.8612 (±5e−4)	0.2059 (±4e−4)
	Bayesian Output	0.4382 (±2e−3)	0.8714 (±5e−4)	0.2189 (±6e−4)	0.3702 (±1e−3)	0.8572 (±5e−4)	0.2246 (±6e−4)
	Bayesian Hidden+Output	0.4492 (±1e−3)	0.8751 (±5e−4)	0.2045 (±5e−4)	0.3893 (±1e−3)	0.8607 (±5e−4)	0.2102 (±5e−4)
	Bayesian RNN+Hidden+Output	0.4396 (±2e−3)	0.8673 (±5e−4)	0.2109 (±5e−4)	0.3860 (±2e−3)	0.8542 (±5e−4)	0.2146 (±5e−4)
	Fully Bayesian	0.4354 (±2e−3)	0.8692 (±5e−4)	0.2068 (±5e−4)	0.3829 (±1e−3)	0.8552 (±5e−4)	0.2103 (±5e−4)
eICU	Deterministic Ensemble	0.1951 (±1e−3)	0.7882 (±7e−4)	0.1435 (±3e−4)	0.2196 (±1e−3)	0.7868 (±6e−4)	0.2435 (±5e−4)
	Bayesian Embeddings	0.1996 (±1e−3)	0.7807 (±1e−4)	0.1455 (±4e−4)	0.2244 (±1e−3)	0.7733 (±7e−4)	0.1620 (±4e−4)
	Bayesian Output	0.1738 (±1e−3)	0.7677 (±7e−4)	0.1664 (±3e−4)	0.1942 (±1e−3)	0.7580 (±7e−4)	0.1810 (±4e−4)
	Bayesian Hidden+Output	0.1712 (±1e−3)	0.7801 (±7e−4)	0.1619 (±3e−4)	0.2140 (±1e−3)	0.7817 (±6e−4)	0.1713 (±3e−4)
	Bayesian RNN+Hidden+Output	0.1675 (±1e−3)	0.7791 (±7e−4)	0.1477 (±3e−4)	0.2147 (±1e−3)	0.7809 (±7e−4)	0.1583 (±3e−4)
	Fully Bayesian	0.2004 (±1e−3)	0.7910 (±7e−4)	0.1377 (±3e−4)	0.2280 (±1e−3)	0.7818 (±7e−4)	0.1541 (±4e−4)

able to understand when a model is uncertain about the prediction it is making.

Figure 2 visualizes the means versus standard deviations of the predictive uncertainty distributions for the deterministic ensemble on all validation set examples. In contrast to the variance of a Bernoulli distribution, which is a simple function of the mean, we find that the standard deviations are patient-specific, and thus cannot be determined *a priori*. In Figure 3, we plot the standard deviations and differences between the maximum and minimum predicted probability values for each patient's predictive uncertainty distribution, $p(\lambda|\mathbf{x})$. We find that there is wide variability in predicted probabilities for some patients, and that negative patients have less variability on average.

Optimal Decision Distributions & Statistics. In practice, model uncertainty is important insofar as it can affect the model's decisions. To test this, we optimize the sensitivity-based (*i.e.*, recall-based) decision cost function with respect to the probability threshold for each model in our RNN ensemble separately to achieve a recall

of 70%, and then make optimal decisions for each example with each of the M models. Figure 4 visualizes how model uncertainty in probability space is realized in optimal decision space for two patients in the mortality task. We see that the model uncertainty does indeed extend into the optimal decision space, leading to a set of optimal decisions for a given patient that can be represented as a distribution over the optimal decision. Furthermore, the decision distribution's variance can be quite high, and knowing when this is the case is important in order to avoid the cost of any incorrect decisions made by the system due to lack of precise knowledge about the correct predictive distribution $p(y|\mathbf{x})$ (*i.e.*, the correct level of data uncertainty).

Figure 5 examines the distribution of maximum predicted probabilities over the CCS classes, along with the distribution of predicted classes associated with the maximum probabilities. Similar to the binary mortality task, this demonstrates the presence of disagreement due to model uncertainty in the multiclass clinical setting.
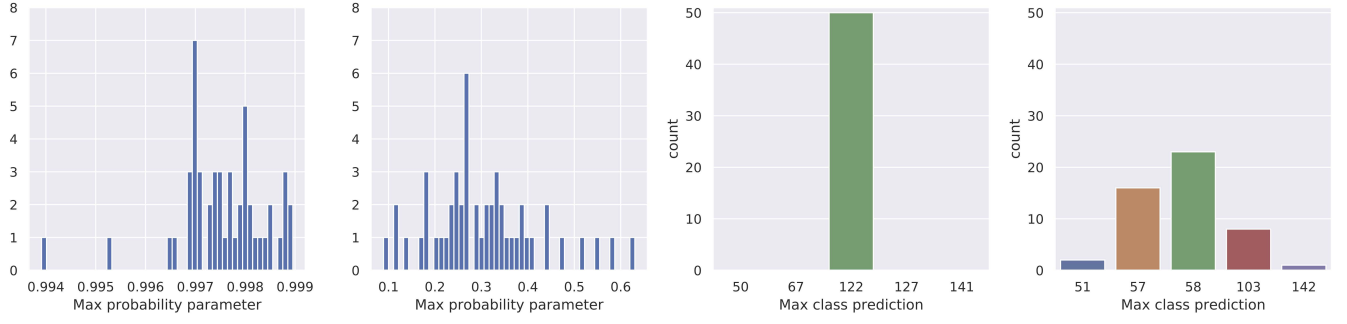
Figure 5: *Left two:* **A set of distributions for the maximum predicted probability from our deterministic ensemble for two patients in the validation dataset on the multiclass CCS diagnosis code task. Note the difference in x-axis scales.** *Right two:* **The corresponding distributions of classes associated with the max probabilities. Similar to the mortality task, for one patient, the ensemble is relatively certain about the predicted class, while for the other patient, there is a larger amount of disagreement.**
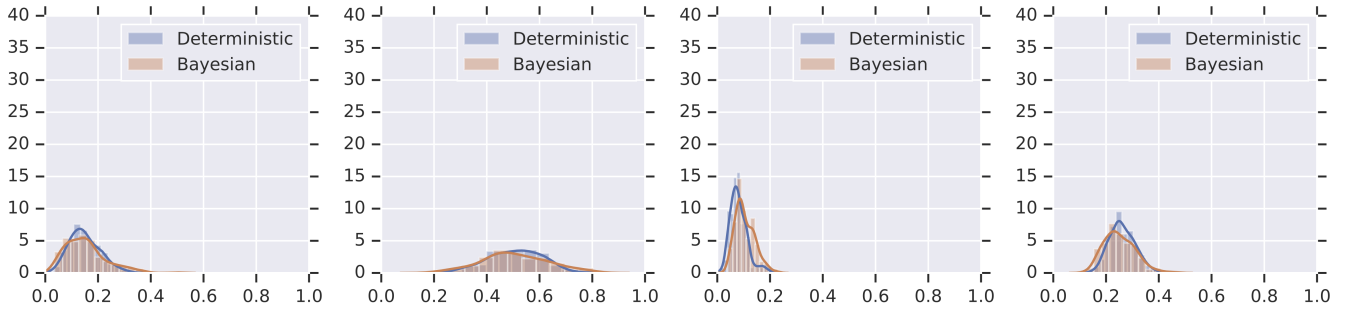

Figure 6: Predictive uncertainty distributions of both the RNN with Bayesian embeddings and the deterministic RNN ensemble for individual patients. We find that the Bayesian model is qualitatively able to capture model uncertainty that aligns with that of the ensemble, while using a considerably smaller number of parameters.

4.2 Comparison: Variants of Bayesian RNNs and Deterministic RNN Ensembles

A natural question in practice when employing the Bayesian approach is: which part of the model should capture model uncertainty? To answer this question, we study Bayesian RNNs under a variety of priors:

- **Bayesian Embeddings** A Bayesian RNN in which the embedding parameters are stochastic, and all other parameters are deterministic.

- **Bayesian Output** A Bayesian RNN in which the output layer parameters are stochastic, and all other parameters are deterministic.

- **Bayesian Hidden+Output** A Bayesian RNN in which the hidden and output layer parameters are stochastic, and all other parameters are deterministic.

- **Bayesian RNN+Hidden+Output** A Bayesian RNN in which the LSTM, hidden, and output layer parameters are stochastic, and all other parameters are deterministic.

- **Fully Bayesian** A Bayesian RNN in which all parameters are stochastic.

Table 2 displays AUC-PR, AUC-ROC, and negative log-likelihood (NLL) metrics over marginalized predictions for each of the Bayesian RNN models and the deterministic RNN ensemble on the MIMIC-III and eICU mortality tasks. We find that the Bayesian Embeddings RNN model outperforms all other Bayesian variants and slightly outperforms the deterministic RNN ensemble in terms of AUC-PR for MIMIC-III, and that the fully-Bayesian RNN outperforms the other models on the eICU dataset. Additionally, all of the Bayesian variants are either comparable or outperform the deterministic ensemble in terms of held-out NLL on both datasets.

Figure 6 visualizes the predictive distributions of both the Bayesian RNN with Bayesian embeddings, and the deterministic RNN ensemble for four individual patients on the MIMIC-III mortality task. The aim is to determine whether the two models are capturing the same distribution over functions insofar as they each produce the same distribution $p(\lambda|\mathbf{x})$ for a given patient $\mathbf{x}$. We find that the Bayesian model qualitatively captures model uncertainty that aligns with that of the deterministic ensemble. Overall, the Bayesian Embeddings RNN, compared to the deterministic RNN ensemble, demonstrates slightly improved predictive performance and qualitatively similar model uncertainty.
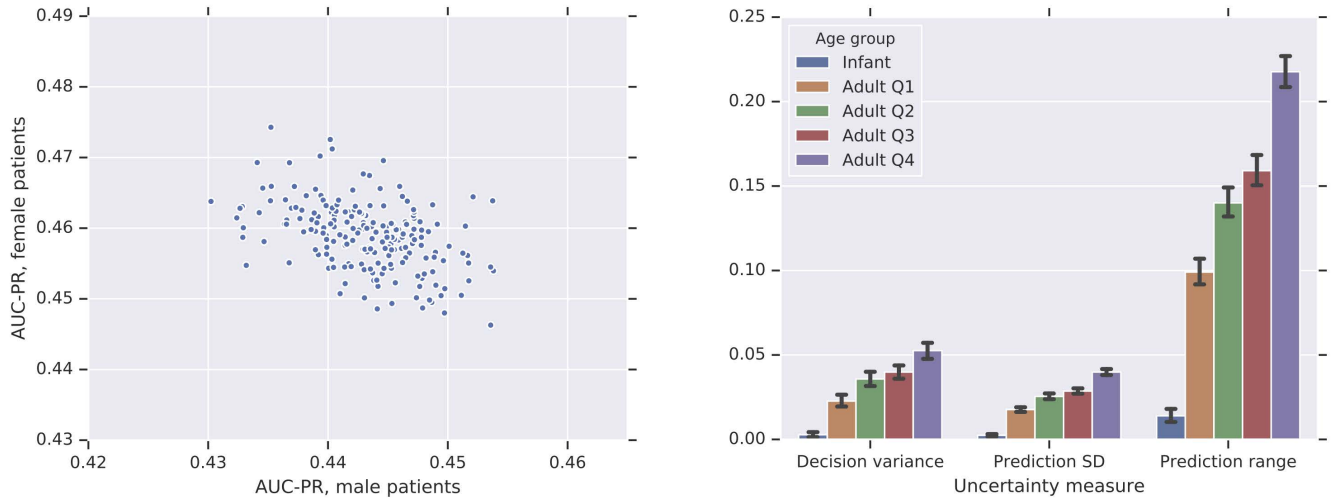
Figure 7: Left: Model performance comparison on male vs. female patients. Each point represents stratified AUC-PR for a single model from the deterministic ensemble. Correlation coefficient $r = -0.442$. Right: Summary of uncertainty measures within each age subgroup, using the Bayesian Embeddings RNN. On all measures, uncertainty increases monotonically with age. This corresponds to an increase in mortality rate with age, as positive cases are more uncertain on average.

Our Bayesian models achieve strong performance while only requiring training of a single model (7.22 million parameters in the MIMIC-III Bayesian Embeddings RNN), versus M models in the deterministic RNN ensemble ($M \times 6.16$ million parameters), as well as only requiring a single model at prediction time. In the deterministic ensemble case, we must choose the number of models M *a priori*, where M affects the level of detail we can expect to obtain in our predictive distributions. With a Bayesian model, we can choose the number of samples M to draw at prediction time, dynamically adjusting it as we see fit. With considerably less computational resources required, using Bayesian RNNs can be a more efficient approach, making it an attractive choice for deployment in clinical practice.

4.3 Patient Subgroup Analysis

We next turn to an exploration of the effects of model uncertainty across patient subgroups. We split validation set encounters into subgroups by demographic characteristics, namely patient gender (3089 male vs. 2548 female) and age (adults divided into quartiles of 1216, with a separate fifth group of 773 neonates). For this analysis, we focus on the deterministic RNN ensemble described in Section 4.1, as the Bayesian models sample $M = 200$ weights for each prediction separately rather than globally for repeated usage across the complete validation set. For each model in the ensemble, we compute validation set performance metrics separately over each subgroup and then compute the correlation between these metrics over all models in the ensemble to evaluate whether the ensemble models tend to specialize to one or more subgroups at the cost of performance on others. We find some evidence of this phenomenon: for example, AUC-PR for male patients is negatively correlated with AUC-PR for female patients (Pearson's $r = -0.442$, see Figure 7),

and AUC-PR for the oldest quartile of adult patients is somewhat negatively correlated with AUC-PR for other adults or for neonates (Pearson's r between -0.18 and -0.37).

We also compare uncertainty metrics across subgroups, including standard deviation and range of the predictive uncertainty distributions, and variance of the optimal decision distributions for patients in each subgroup. For this analysis, we examine both the deterministic RNN ensemble and the best Bayesian model, the RNN using Bayesian embeddings. In both cases, we find that all metrics are correlated with subgroup label prevalence: both uncertainty and mortality rate increase monotonically across age groups (Figure 7), and both are slightly higher in women than in men. These findings imply that random model variation during training may actually cause unintentional harm to certain patient populations, which may not be reflected in aggregate performance.

4.4 Embedding Uncertainty Analysis

Another motivation for model uncertainty lies in understanding which feature values are most responsible for the variance of the predictive uncertainty distribution. Our RNN with Bayesian embeddings model is particularly well suited for this task in that the uncertainty in embedding space directly corresponds to the predictive uncertainty distribution and represents uncertainty associated with the discrete feature values. Understanding model uncertainty associated with features can provide some level of interpretability by allowing us to recognize particularly difficult examples and understand which feature values are leading to the disagreement amongst models. Additionally, it provides a means of determining the types of patient examples that could be beneficial to add to the training dataset for future updates to the model.

Table 3: Top and bottom 10 words in free-text clinical notes with the highest and lowest model uncertainty based on the entropy of each word's associated Bayesian embedding distribution, along with total counts in the training set.

LOWEST UNCERTAINTY		
WORD	ENTROPY	COUNT
THE	-82.5444	41803
AND	-80.6054	42812
OF	-80.2735	43191
NO	-79.8993	43420
TRACING	-78.5987	32181
IS	-78.5552	42560
TO	-77.6408	42365
FOR	-76.8005	42972
WITH	-75.3513	42819
IN	-72.8005	42144

HIGHEST UNCERTAINTY		
WORD	ENTROPY	COUNT
24PM	-16.0789	336
LABWORK	-16.0749	272
COLONIAL	-16.0689	198
ZOYSN	-16.0601	269
HT	-16.0522	515
TXCF	-15.9982	112
ARRANGEMENTS	-15.9794	407
PARVUS	-15.9773	132
NAS	-15.9163	251
ANESTHESIOLOGIST	-15.8796	220

For this analysis, we focus on the free-text clinical notes found in the EHR. For each word in the notes vocabulary, we have an associated embeddings distribution formulated as a multivariate normal distribution. We rank each word by its level of model uncertainty, which we measure in this case by the entropy of its embedding distribution. Table 3 lists the top and bottom ten words, along with each word's count in the training dataset. We find, in general, that common words, both subjectively and based on prevalence counts, have lower entropy and thus limited model uncertainty, while rarer words have higher entropy levels, which corresponds to higher model uncertainty. However, there is a nonlinear relationship between prevalence and entropy, which can be seen, for example, with the word "tracing", which has approximately a 25% lower count than the other nine words in the bottom ten words, yet has the fifth lowest entropy. This provides some evidence that the model uncertainty is context-specific.

We additionally measure the correlation between entropy and word frequency as visualized in Figure 8. We find further confirmation that rarer words are generally associated with higher model

uncertainty, but that there is a nonlinear relationship between the two entities.
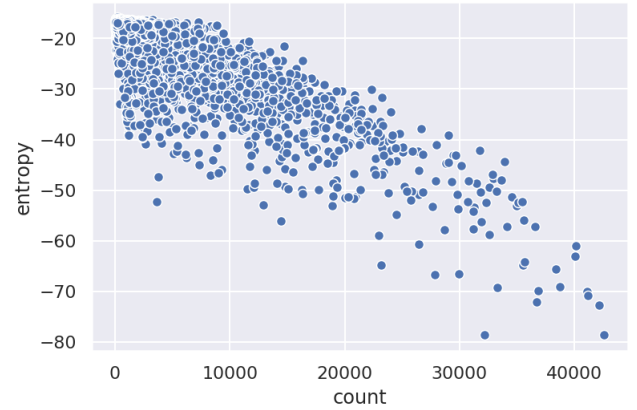

Figure 8: Correlation between the entropy of the Bayesian embedding distributions for free-text clinical notes and the associated word frequency. We find that rarer words are associated with higher model uncertainty, with a non-trivial level of variance at a given frequency.

5 CONCLUSION

In this work, we demonstrated the need for capturing model uncertainty in medicine and examined methods to do so. Our experiments showed multiple findings. For example, an ensemble of deterministic RNNs captured individualized uncertainty that led to high predictive disagreement for some patients, all while the models each maintained nearly equivalent clinically-relevant dataset-level metrics. Furthermore, this disagreement propagated forward as disagreement over the optimal decision for a given patient. Significant variability in patient-specific predictions and decisions can be an indicator of when a model decision is likely to be brittle, and it provides an opportunity to identify and collect additional data that could reduce the level of model uncertainty. As another example, we found that models need only be uncertain around the embeddings for competitive performance, as seen by the RNN with Bayesian embeddings. This provided an additional benefit of enabling the ability to determine the level of model uncertainty associated with individual feature values, allowing for some level of interpretability. Furthermore, using model uncertainty methods, we examined patterns in uncertainty across patient subgroups, showing that models can exhibit higher levels of uncertainty for certain groups.

Future work includes designing more specific and clinically-relevant decision cost functions based on both quantified medical ethics [10] and monetary axes; making optimal decisions in light of both data and model uncertainty; and exploring methods to reduce model uncertainty at both training and prediction time.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 265–283.

[2] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning* (corrected 8th printing 2009 ed.). Springer, New York.

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight Uncertainty in Neural Networks. *arXiv.org* (May 2015). http://arxiv.org/abs/1505.05424v2

[4] Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. 2018. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *Advances in Neural Information Processing Systems*. 4547–4557.

[5] National Research Council et al. 2011. *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease.* National Academies Press.

[6] Michael W Dusenberry, Charles K Brown, and Kori L Brewer. 2017. Artificial neural networks: Predicting head CT findings in elderly patients presenting with minor head injury after a fall. *The American journal of emergency medicine* 35, 2 (Feb. 2017), 260–267. https://doi.org/10.1016/j.ajem.2016.10.065

[7] Meire Fortunato, Charles Blundell, and Oriol Vinyals. 2017. Bayesian Recurrent Neural Networks. *arXiv.org* (April 2017). http://arxiv.org/abs/1704.02798v3

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

[9] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. 2018. Neural processes. *arXiv preprint arXiv:1807.01622* (2018).

[10] Raanan Gillon. 1994. Medical ethics: four principles plus attention to scope. *Bmj* 309, 6948 (1994), 184.

[11] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D. Sculley. 2017. Google Vizier: A Service for Black-Box Optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*. ACM Press, Halifax, NS, Canada, 1487–1495. https://doi.org/10.1145/3097983.3098043

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning (ICML)*, Vol. cs.LG. Cornell University Library. http://arxiv.org/abs/1706.04599v2

[13] Danijar Hafner, Dustin Tran, Alex Irpan, Timothy Lillicrap, and James Davidson. 2018. Reliable uncertainty estimates in deep neural networks using noise contrastive priors. *arXiv preprint arXiv:1807.09289* (2018).

[14] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2017. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771* (2017).

[15] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data* 3 (May 2016), 160035. https://doi.org/10.1038/sdata.2016.35

[16] Alex Kendall and Yarin Gal. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?. In *Advances in Neural Information Processing Systems*. http://arxiv.org/abs/1703.04977

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Frank H. Knight. 1957. *Risk, Uncertainty and Profit.* New York, Kelley & Millman. https://mises.org/sites/default/files/Risk,%20Uncertainty,%20and%20Profit_4.pdf

[19] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. 2017. Automatic differentiation variational inference. *The Journal of Machine Learning Research* 18, 1 (2017), 430–474.

[20] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Advances in Neural Information Processing Systems*, Vol. stat.ML. http://arxiv.org/abs/1612.01474v3

[21] Christos Louizos and Max Welling. 2017. Multiplicative normalizing flows for variational bayesian neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2218–2227.

[22] Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*. 7047–7058.

[23] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *AAAI Conference on Artificial Intelligence*, Vol. 2015. 2901–2907. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4410090/pdf/nihms679964.pdf

[24] Jeremy Nixon, Mike Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring Calibration in Deep Learning. *arXiv:1904.01685 [cs, stat]* (April 2019). http://arxiv.org/abs/1904.01685

[25] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data* 5 (2018).

[26] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. 2018. Scalable and Accurate Deep Learning with Electronic Health Records. *Nature Partner Journals: Digital Medicine* 1, 1 (May 2018), 18. https://doi.org/10.1038/s41746-018-0029-1

[27] Teri A Reynolds. 2013. A Tunisian, a Canadian, and an American walk into a bar (sustaining mild head injury). *Annals of Emergency Medicine* 61, 5 (2013), 528–529. https://doi.org/10.1016/j.annemergmed.2012.12.006 tex.date-modified: 2016-06-13 16:42:01 +0000 tex.publisher: Elsevier.

[28] Juergen Schmidhuber and Sepp Hochreiter. 1997. Long Short-Term Memory. *Neural computation* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/doi.org/10.1162/neco.1997.9.8.1735

[29] Marion Smits. 2005. External Validation of the Canadian CT Head Rule and the New Orleans Criteria for CT Scanning in Patients With Minor Head Injury. *JAMA* 294, 12 (Sept. 2005), 1519. https://doi.org/10.1001/jama.294.12.1519

[30] Ian G Stiell, Catherine M Clement, Brian H Rowe, Michael J Schull, Robert Brison, Daniel Cass, Mary A Eisenhauer, R Douglas McKnight, Glen Bandiera, Brian Holroyd, and others. 2005. Comparison of the Canadian CT Head Rule and the New Orleans Criteria in patients with minor head injury. *JAMA* 294, 12 (2005), 1511–1518. https://doi.org/10.1001/jama.294.12.1511 tex.date-modified: 2016-06-13 16:51:15 +0000 tex.publisher: American Medical Association.

[31] Ian G Stiell, George A Wells, Katherine Vandemheen, Catherine Clement, Howard Lesiuk, Andreas Laupacis, R Douglas McKnight, Richard Verbeek, Robert Brison, Daniel Cass, Mary A Eisenhauer, Gary H Greenberg, and James Worthington. 2001. The Canadian CT Head Rule for patients with minor head injury. *The Lancet* 357, 9266 (May 2001), 1391–1396. https://doi.org/10.1016/S0140-6736(00)04561-X

[32] Dustin Tran, Michael W. Dusenberry, Mark van der Wilk, and Danijar Hafner. 2018. Bayesian Layers: A Module for Neural Network Uncertainty. *arXiv:1812.03973 [cs, stat]* (Dec. 2018). http://arxiv.org/abs/1812.03973

[33] Yeming Wen, Dustin Tran, and Jimmy Ba. 2020. BatchEnsemble: an Alternative Approach to Efficient Ensemble and Lifelong Learning. *arXiv:2002.06715 [cs, stat]* (Feb. 2020). http://arxiv.org/abs/2002.06715 arXiv: 2002.06715.

[34] Andrew Gordon Wilson. 2019. *The case for Bayesian deep learning.* Technical Report. NYU Courant Institute of Mathematical Sciences and Center for Data Science. https://cims.nyu.edu/~andrewgw/caseforbdl.pdf

[35] Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. 2018. RAIM: Recurrent Attentive and Intensive Model of Multimodal Patient Monitoring Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2565–2573.