Analyzing the Role of Model Uncertainty for Electronic Health Records

Michael W. Dusenberry[1*] **Dustin Tran**[1] **Edward Choi**[2] **Jonas Kemp**[2]
Jeremy Nixon[1] **Ghassen Jerfel**[2,3] **Katherine Heller**[2] **Andrew M. Dai**[2]
[1]Google Brain [2]Google Health [3]Duke University

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.

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 [1], and there can be severe costs for incorrect predictions and decisions due to the risk to human life and the associated ethical concerns [2].

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 [3–6]. These works, while focusing on capturing the data uncertainty, do not address the *model* uncertainty that is inherent in fitting deep neural networks. 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 [7–9], Monte Carlo dropout [10], ensembles [11], and function priors [12–14]. Model uncertainty methods raise several questions:

- How do the realized functions, such as individual models in the ensembles approach, compare in terms of metric performance such as AUC-PR/ROC, log-likelihood, or calibration?
- What is the effect of uncertainty on predictions across patient subgroups, such as by race, gender, age, or length of stay?

[*]Work done as a Google AI Resident.

Machine Learning for Health (ML4H) at NeurIPS 2019 - Extended Abstract

- Which feature values are responsible for the highest 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 clinical dataset [15], 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, 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

Model uncertainty Model uncertainty can be viewed as uncertainty in the correct values of the parameters for the predictive outcome distribution due to not knowing the true function. For binary tasks, this equates to a distribution of plausible probability values for a Bernoulli distribution,

$$y \sim \text{Bernoulli}(\lambda), \quad \lambda \sim p(\lambda|\mathbf{x}, \mathbf{w}), \tag{1}$$

where there is a distribution over the Bernoulli parameter λ for a given example, conditioned on the inputs $\mathbf{x}$ and parameters $\mathbf{w}$. This represents uncertainty in the true outcome distribution due to uncertainty in function space, and we refer to it as the *predictive uncertainty distribution*.

3 Medical Uncertainty

Clinical Tasks We show results on two clinical tasks: in-patient mortality prediction, and single-level Clinical Classifications Software (CCS) multiclass diagnosis prediction. Similar to Rajkomar et al. [4], we train deep recurrent neural networks (RNNs) on MIMIC-III [15], an EHR dataset collected from 46,520 patients admitted to intensive care units (ICUs) at Beth Israel Deaconess Medical Center, where 9,974 expired during the encounter (*i.e.,* 1:4 ratio between positive and negative samples). Our model embeds and aggregates a patient's time-series features (*e.g.* medications, lab measures) and global features (*e.g.* gender, age), feeds them to one or more long short-term memory (LSTM) layers [16], and follows that with hidden and output affine layers.

Predictive Uncertainty Distributions To quantify model uncertainty for our mortality prediction task, 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 [17], and then train M replicas of the best model. Only the random seed differs between the replicas. At prediction time, we make predictions with all M models for each patient. 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 RNNs to use factorized weight posteriors $q(\mathbf{w}|\boldsymbol{\theta}) = \prod_i q(\mathbf{w}_i|\boldsymbol{\theta}_i)$,where weight tensors $\mathbf{w}_i$ in the models are represented by normal distributions with learnable mean and diagonal covariance parameters represented as $\boldsymbol{\theta}_i$. Normal distributions with zero mean and tunable standard deviation are used as weight priors. We train our models by minimizing the Kullback-Leibler (KL) divergence

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \text{KL}[q(\mathbf{w}|\boldsymbol{\theta}) \,\|\, p(\mathbf{w}|\mathbf{y}, \mathbf{X})] \\ &\propto \text{KL}[q(\mathbf{w}|\boldsymbol{\theta}) \,\|\, p(\mathbf{w})] - \mathbb{E}_q \left[\ln p(\mathbf{y}|\mathbf{X}, \mathbf{w})\right] \end{aligned} \tag{2}$$

between the approximate weight posterior and the true, but unknown posterior, which overall equates to minimizing an expectation over the negative log likelihood term plus a KL regularization term.

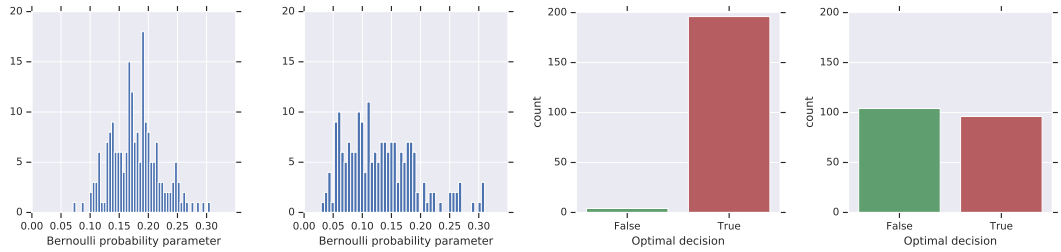

Figure 1: Left two: A set of predictive uncertainty distributions from our deterministic ensemble for two patients in the validation dataset on the mortality task. **Right two:** The corresponding optimal decision distributions. For some patients, the ensemble is relatively certain about the optimal decision, while for other patients, there is a large amount of disagreement.

Optimal Decisions Via Sensitivity Requirements While predicted probabilities of a clinical outcome are important, the key desire for clinical practice is to make an optimal decision. Given a set of potential outcomes y_k, a set of conditional probabilities $p(y_k|\mathbf{x})$ for the given outcomes, and the associated costs L_{kj} of correctly and incorrectly predicting the correct outcome, an optimal decision can be determined by minimizing the expected decision loss

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(y_k|\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \tag{3}$$

where $\mathcal{R}_j$ is the decision region for assigning example $\mathbf{x}$ to class j, and $p(\mathbf{x})$ is the density of $\mathbf{x}$ [18].

We make use of a clinically relevant cost: sensitivity requirements, where the goal is to maximize the precision of a given model on a validation dataset with respect to a probability threshold t_m, subject to a lower bound constraint on the sensitivity. For each of the M models in an ensemble, or M weight samples from a Bayesian model, we can optimize our decision cost function with respect to t_m. Then, for each example, we can make an optimal decision with each model, yielding a set of M optimal decisions. The optimal decision d for a given example can then be represented as a random variable

$$d \sim \text{Bernoulli}(\phi), \quad \phi = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}(\lambda_m \geq t_m)), \tag{4}$$

where $\lambda_m \geq t_m$ is the decision function for model m, and ϕ is the percentage of model agreement.

4 Experiments

Clinical Metrics For our clinical tasks, we measure area under the precision-recall curve (AUC-PR), area under the receiver operating characteristic curve (AUC-ROC), top-5 recall, top-5 precision, top-5 F1, held-out log-likelihood, expected calibration error (ECE) [19], and adaptive calibration error (ACE) [20]. Table 1 shows the mean and standard deviation of the metrics for the models in our deterministic ensemble. For the ensemble, we find that the models are well-calibrated and nearly equivalent in terms of performance. It is highly likely that any one could have been selected in practice if we were only using one model for our clinical problem.

Predictive Uncertainty Distributions & Statistics Figure 1-left visualizes the predictive uncertainty distributions for two different patients. We find that there is a wide variability in predicted Bernoulli probabilities for some patients (with spreads as high as 57.5%), despite the models being nearly identical in terms of dataset-level metrics.

Optimal Decision Distributions & Statistics Figure 1-right visualizes how model uncertainty in probability space is realized in optimal decision space for two different patients given a sensitivity constraint of at least 70%. We see that the model uncertainty extends into the optimal decision space, converting the optimal decision into a random variable. Furthermore, the decision distribution's variance can be quite high, and knowing when this is the case is important in order to identify when a model-assisted clinical decision is likely to be brittle. Using a single model or an average over models masks this information.

Table 1: Metric statistics for the binary mortality and multiclass CCS tasks across $M = 200$ models in the deterministic RNN ensemble. The models are nearly identical in terms of dataset-level performance across both tasks, despite having high disagreement for some patients.

Task	Metric	Validation	Test
Mortality	AUC-PR	0.4496 (0.0025)	0.3886 (0.0059)
	AUC-ROC	0.8753 (0.0019)	0.8623 (0.0031)
	ECE	0.0176 (0.0040)	0.0162 (0.0043)
	ACE	0.0210 (0.0042)	0.0233 (0.0057)
CCS	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)
	Log-likelihood	-5.1040 (0.0075)	-5.1081 (0.0083)
	ECE	0.0446 (0.0072)	0.0499 (0.0082)
	ACE	4.2189e-3 (7.3111e-8)	4.2191e-3 (7.6136e-8)

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

Model	Val. AUC-PR	Val. AUC-ROC	Val. log-likelihood	Test AUC-PR	Test AUC-ROC	Test log-likelihood
Deterministic	0.4564	0.8774	-0.7113	0.3921	0.8643	-0.7148
Ensemble	$(\pm 1e{-3})$	$(\pm 5e{-4})$	$(\pm 7e-5)$	$(\pm 1e{-3})$	$(\pm 5e{-4})$	$(\pm 7e-5)$
Bayesian	0.4580	0.8776	-0.7152	0.3977	0.8612	-0.7186
Embeddings	$(\pm 1e-3)$	$(\pm 4e-4)$	$(\pm 7e-5)$	$(\pm 2e-3)$	$(\pm 5e-4)$	$(\pm 7e-5)$
Bayesian	0.4382	0.8714	-0.7058	0.3702	0.8572	-0.7087
Output	$(\pm 2e{-3})$	$(\pm 5e{-4})$	$(\pm 7e-5)$	$(\pm 1e-3)$	$(\pm 5e-4)$	$(\pm 7e-5)$
Bayesian	0.4492	0.8751	-0.7118	0.3893	0.8607	-0.7149
Hidden+Output	$(\pm 1e-3)$	$(\pm 5e-4)$	$(\pm 7e-5)$	$(\pm 1e-3)$	$(\pm 5e-4)$	$(\pm 7e-5)$
Bayesian	0.4396	0.8673	-0.7097	0.3860	0.8542	-0.7125
RNN+Hidden+Output	$(\pm 2e-3)$	$(\pm 5e-4)$	$(\pm 7e-5)$	$(\pm 2e-3)$	$(\pm 5e-4)$	$(\pm 7e-5)$
Fully	0.4354	0.8692	-0.7100	0.3829	0.8552	-0.7133
Bayesian	$(\pm 2e-3)$	$(\pm 5e-4)$	$(\pm 6e-5)$	$(\pm 1e-3)$	$(\pm 5e-4)$	$(\pm 6e-5)$

Bayesian Variants A natural question in practice is where precisely to be uncertain in the model. To do so, we convert our deterministic RNN to a Bayesian RNN, and study it under a variety of priors. Table 2 displays the metrics over marginalized predictions for the each of the models. We find that the Bayesian Embeddings RNN model outperforms all other Bayesian variants and slightly outperforms the deterministic ensemble in terms of AUC-PR. Additionally, all of the Bayesian variants are either comparable to or outperform the deterministic ensemble in terms of log-likelihood. Since our Bayesian models achieve strong performance while only requiring training of a single model (7.22M parameters in the Bayesian Embeddings RNN), versus M models in the deterministic ensemble (200 * 6.16M parameters), using Bayesian RNNs can be a more efficient approach.

Patient Subgroup Analysis For this analysis, we compare uncertainty metrics across subgroups, including the standard deviation and range of the predictive uncertainty distributions, and the variance of the optimal decision distributions for patients in each subgroup. We find that model uncertainty is correlated with subgroup label prevalence: both uncertainty and mortality rate increase monotonically across age groups (Figure A.3), 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.

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. For this analysis, we focus on the free-text clinical notes found in the EHR, and examine our Bayesian embeddings RNN. For each word in the notes vocabulary, we measure the entropy of its associated embedding distribution, and then rank them. We find that common words have generally have low entropy and thus limited model uncertainty, while rarer words generally have higher entropy levels, corresponding to higher model uncertainty (Table A.2). Understanding model uncertainty associated with features can allow us to recognize particularly difficult examples and understand which feature values are leading to the difficulties. Additionally, it provides a means of determining the types of examples that could be beneficial to add to the training dataset for future updates to the model.

References

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

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

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

[4] 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. Scalable and accurate deep learning with electronic health records. *Nature Partner Journals: Digital Medicine*, 1(1):18, May 2018. ISSN 2398-6352. doi: 10.1038/s41746-018-0029-1.

[5] Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. 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*, pages 2565–2573. ACM, 2018.

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

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

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

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

[10] 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, 2016.

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

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

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

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

[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. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL `http://www.nature.com/articles/sdata201635`.

[16] Juergen Schmidhuber and Sepp Hochreiter. Long short-term memory. *Neural computation*, 9 (8):1735–1780, November 1997. doi: doi.org/10.1162/neco.1997.9.8.1735.

[17] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D. Sculley. 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*, pages 1487–1495, Halifax, NS, Canada, 2017. ACM Press. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098043. URL `https://ai.google/research/pubs/pub46180.pdf`.

[18] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, corrected 8th printing 2009 edition, 2006. ISBN 978-0-387-31073-2.

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

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

[21] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5, 2018.

[22] Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine*, 169(12):866, December 2018. ISSN 0003-4819. doi: 10.7326/M18-1990. URL `http://annals.org/article.aspx?doi=10.7326/M18-1990`.

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

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

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

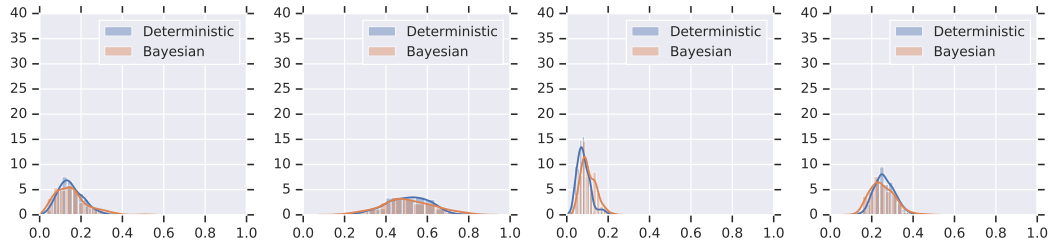

Figure A.1: Predictive uncertainty distributions of both the RNN with Bayesian embeddings and the deterministic RNN ensemble for individual patients. We find that the Bayesian model is qualitatively able to capture model uncertainty that is quite similar to that of the ensemble.

Table A.1: Metrics for marginalized predictions on the eICU mortality task given $M = 200$ models in the deterministic RNN ensemble, and $M = 200$ samples from each of the Bayesian RNN models. Confidence intervals are computed via validation and test set bootstrapping with 1000 bootstrap sets.

Data	Model	Val. AUC-PR	Val. AUC-ROC	Val. NLL	Test AUC-PR	Test AUC-ROC	Test NLL
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 (±e−)	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)

A Appendix

A.1 Bayesian RNNs

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.

Figure A.1 visualizes the predictive distributions of both the RNN with Bayesian embeddings and the deterministic RNN ensemble for four individual patients. We find that the Bayesian model is qualitatively able to capture model uncertainty that is quite similar to that of the deterministic ensemble.

A.2 eICU Collaborative Research Database Dataset

In order to demonstrate that our findings generalize, we additionally experiment with the eICU Collaborative Research Database (eICU) dataset [21], another publicly available EHR dataset. Table A.1 displays the metrics over marginalized predictions for each of the Bayesian RNN models and the deterministic RNN ensemble on the eICU in-patient mortality prediction task. We find that the fully-Bayesian RNN model outperforms all other models, including the deterministic deep RNN

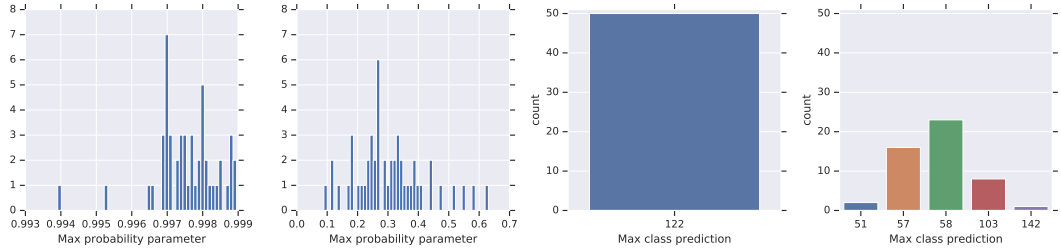

Figure A.2: Left two: A set of distributions for the maximum predicted probability from our deterministic ensemble for two patients in the validation dataset on the mortality task. Note the difference in x-axis scales. **Right two:** The corresponding distributions of classes associated with the max probabilities. Similar to the mortality task, for some patients, such as the one on the left, the ensemble is relatively certain about the predicted class (completely certain in this case), while for other patients, such as the one on the right, there is a larger amount of model uncertainty.

ensemble, on both the validation and test datasets. Similar to the the MIMIC-III experiment, the Bayesian Embeddings RNN also outperforms the deterministic deep RNN ensemble and the other remaining Bayesian variants.

A.3 CCS Multiclass Task Analysis - Extended

Figure A.2 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 model uncertainty in the multiclass clinical setting.

A.4 Patient Subgroup Analysis - Extended

In addition to the analysis of model uncertainty across subgroups, we compare the performance and uncertainty of models from the ensemble across patient subgroups. We split validation set encounters into subgroups by demographic characteristics, namely patient gender (3089 male vs. 2548 female) or 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 separately for each prediction rather than globally for 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 A.3), 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).

A.5 Embedding Uncertainty Analysis - Extended

Table A.2 lists the top and bottom 10 words, along with their frequency in the training dataset. We find that common words, both subjectively and based on prevalence counts, have low entropy and thus limited model uncertainty, while rarer words have higher entropy levels, which corresponds to higher model uncertainty. We additionally measure the correlation between entropy and word frequency as visualized in Figure A.4.

A.6 Additional Metrics and Statistics

Predictive Uncertainty Distributions Figure A.5 visualizes the predictive uncertainty distribution for a single patient. We find that there is a wide variability in predicted Bernoulli probabilities for some patients (with spreads as high as 57.5%), despite the models being nearly identical in terms of

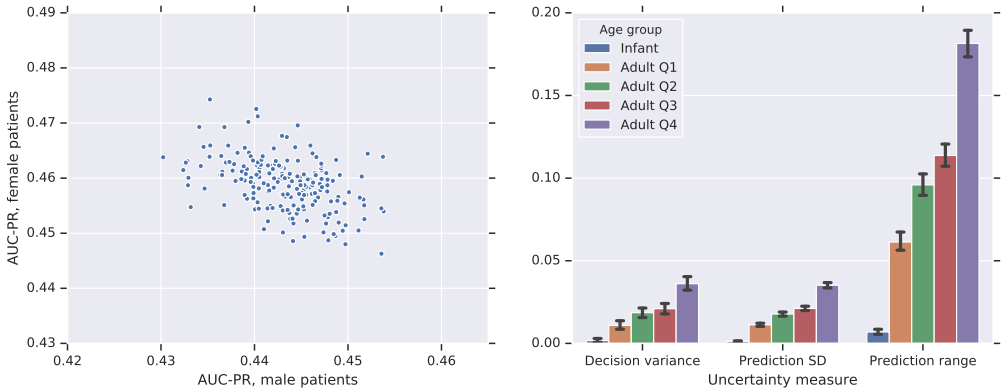

Figure A.3: Left: Model performance comparison on male vs. female patients. Each point represents stratified AUC-PR for a single model. Correlation coefficient $r = -0.442$. **Right:** Summary of uncertainty measures within each age subgroup. On all measures, uncertainty increases monotonically with age. This corresponds to an increase in mortality rate with age, as positive cases are more uncertain on average.

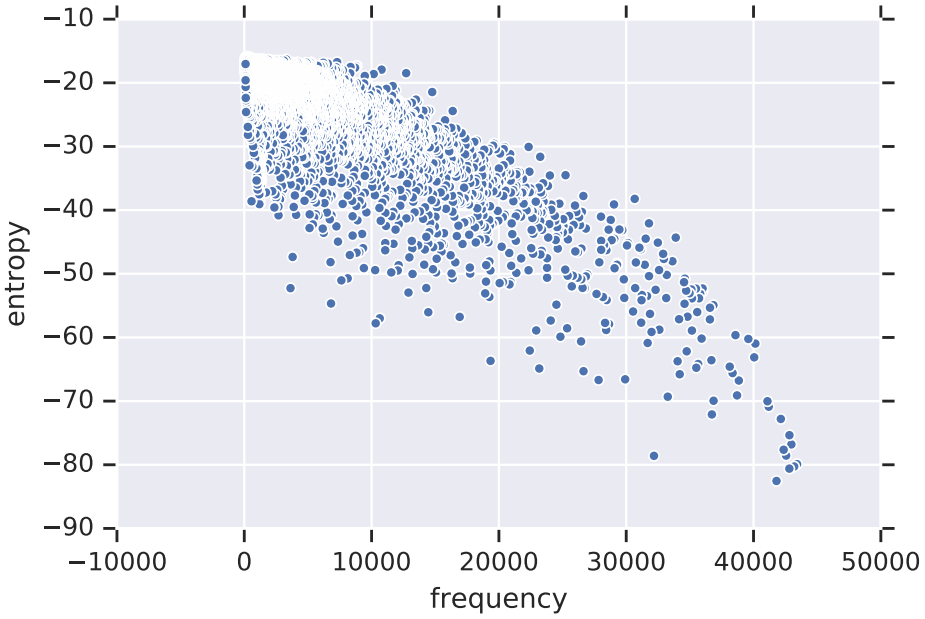

Figure A.4: Correlation between the entropy of the Bayesian embedding distributions for free-text clinical notes and the associated word frequency. We find that rarer words are associated with higher model uncertainty, with some level of variance at a given frequency.

dataset-level metrics. Additionally, Figure A.6 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 A.7, we examine the correlation between held-out log-likelihood and AUC-PR values for models in the deterministic RNN ensemble on the mortality task.

In Figure A.8, we plot the differences between the maximum and minimum predicted probability values for each patient's predictive uncertainty distribution. We find that there is wide variability in predicted probabilities for some patients, and that negative patients have less variability on average.

Table A.2: Top and bottom 10 words in free-text clinical notes based on their associated Bayesian embeddings distribution's entropy, along with their frequency in the training dataset.

Lowest Entropy			Highest Entropy		
Word	Entropy	Frequency	Word	Entropy	Frequency
THE	-82.5445	41803	24PM	-16.0790	336
AND	-80.6055	42812	LABWORK	-16.0750	272
OF	-80.2735	43191	COLONIAL	-16.0690	198
NO	-79.8994	43420	ZOYSN	-16.0601	269
TRACING	-78.5988	32181	HT	-16.0523	515
IS	-78.5553	42560	TXCF	-15.9982	112
TO	-77.6408	42365	ARRANGEMENTS	-15.9795	407
FOR	-76.8005	42972	PARVUS	-15.9773	132
WITH	-75.3513	42819	NAS	-15.9164	251
IN	-72.8006	42144	ANESTHESIOLOGIST	-15.8796	220

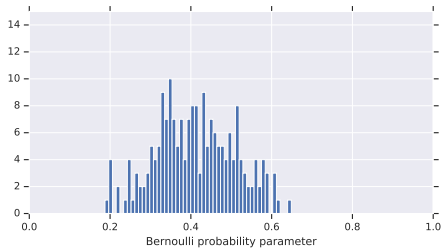

Figure A.5: A histogram of predictions from M RNN models for the probability of mortality for a given patient in the ICU, where the disagreement is due to model uncertainty.

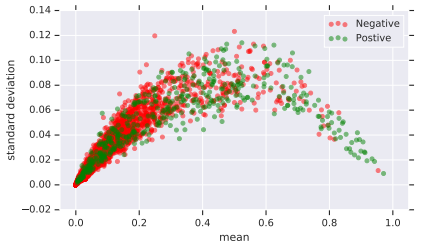

Figure A.6: Mean versus standard deviation of the predictive uncertainty distributions of the deterministic ensemble for positive and negative patients in the validation set.

In Table A.3, we measure the calibration of marginalized predictions of our deterministic RNN ensemble and the Bayesian RNNs.

In Figure A.9, we plot the precision-recall (PR) curves of all M models in our deterministic RNN ensemble across the full dataset along with error bars. We find that the PR curves are nearly identical for all models, and thus it again seems highly likely that any one of the models could have been selected if we were focused on our recall-based decision cost function.

A.7 Additional Training Details

Our RNN model uses the same embedding logic as used in [22] to embed sequential and contextual features. 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 a single probability value. 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. All Bayesian models use stochastic versions of the corresponding deterministic layers, and we make use of the Bayesian Layers setup [23]. More details on the training setup can be found in the code[2].

In terms of hyperparameter optimization, we searched over the hyperparameters listed in Table A.4 for the original deterministic RNN (all others in the ensemble differ only by the random seed) and each of the Bayesian models. Table A.5 lists the final hyperparameters associated with each of the models presented in the paper.

Models were implemented using TensorFlow 2.0 [24], and trained on machines equipped with Nvidia's V100 using the Adam optimizer [25]. To easily shift between the deterministic and Bayesian models, we make use of the Bayesian Layers [23] abstractions. MIMIC-III data were split into train, validation, and test set in 8:1:1 ratio.

[2]Code will be open-sourced.

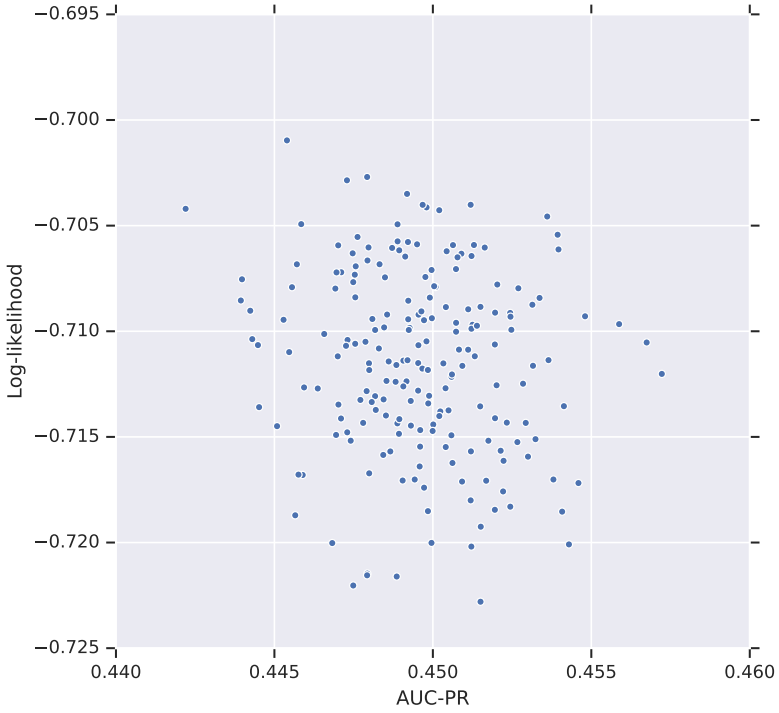

Figure A.7: Validation AUC-PR versus held-out log-likelihood values for the deterministic RNN ensemble on the mortality task. We find that there is no apparent correlation between the two metrics, likely due to the limited differences between the models.

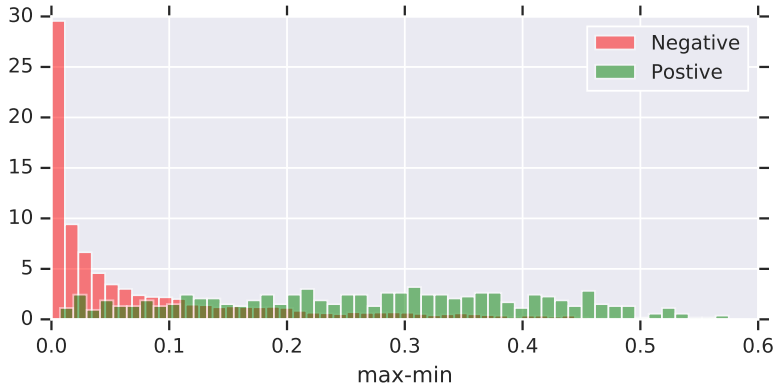

Figure A.8: A histogram of differences between the maximum and minimum predicted probability values for each patient's predictive uncertainty distribution. This shows that there is wide variability in predicted probabilities for some patients, and that negative patients have less variability on average.

Table A.3: Calibration error for marginalized predictions on the mortality task for an average over $M = 200$ models in the deterministic RNN ensemble, and $M = 200$ samples from each of the Bayesian RNN models. We find that marginalization slightly increases the calibration of the deterministic ensemble, and that the Bayesian models are comparably well-calibrated.

Model	Val. ECE	Val. ACE	Test ECE	Test ACE
Deterministic Ensemble	0.0157	0.0191	0.0157	0.0191
Bayesian Embeddings	0.0167	0.0194	0.0163	0.0221
Bayesian Output	0.0263	0.0217	0.0241	0.0279
Bayesian Hidden+Output	0.0194	0.0212	0.0173	0.0240
Bayesian RNN+Hidden+Output	0.0240	0.0228	0.0182	0.0247
Fully Bayesian	0.0226	0.0192	0.0178	0.0197

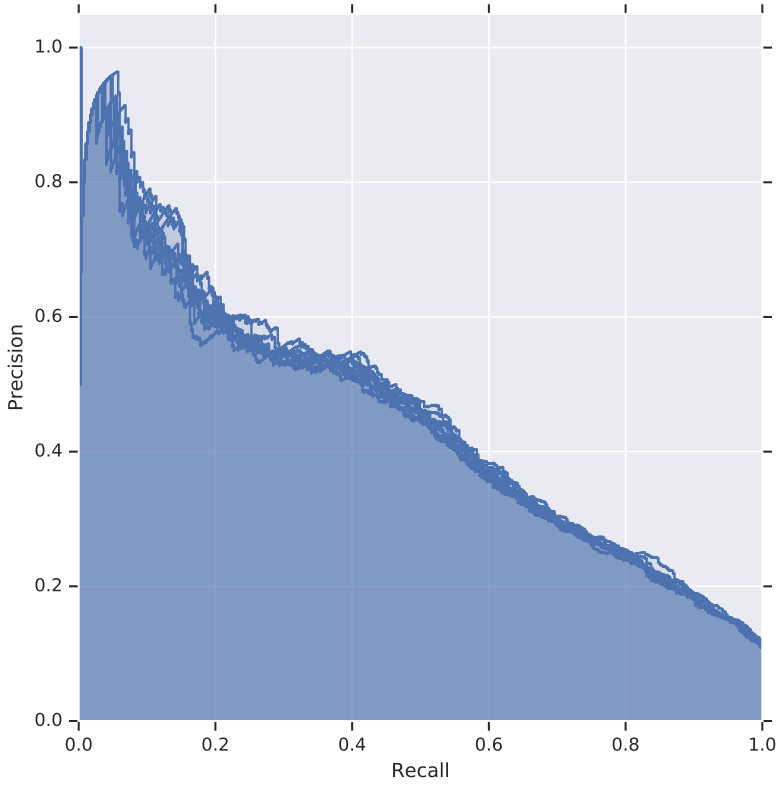

Figure A.9: Precision-recall curves for all $M = 200$ models in the deterministic ensemble. All of the curves are nearly identical, which is in line with the AUC-PR results.

Table A.4: Hyperparameters and their associated search sets or ranges.

Hyperparameter	Range/Set
batch size	{32, 64, 128, 256, 512}
learning rate	[0.00001, 0.1]
KL or regularization annealing steps	[1, 1e6]
prior standard deviation (Bayesian only)	[0.135, 1.0]
Dense embedding dimension	{16, 32, 64, 100, 128, 256, 512}
Embedding dimension multiplier	[0.5, 1.5]
RNN dimension	{16, 32, 64, 128, 256, 512, 1024}
Number of RNN layers	{1, 2, 3}
Hidden affine layer dimension	{0, 16, 32, 64, 128, 256, 512}
Bias uncertainty (Bayesian only)	{True, False}

Table A.5: Model-specific hyperparameter values.

Model	Batch size	Learn. rate	anneal. steps	Prior std. dev.	Dense embed. dim.	Embed. dim. mult.	RNN dim.	Num. RNN layers	Hidden layer dim.	Bias uncert.
Deterministic Ensemble	256	3.035e-4	1	–	32	0.858	1024	1	512	–
Bayesian Embeddings	256	1.238e-3	9.722e+5	0.292	32	0.858	1024	1	512	False
Bayesian Output	256	1.647e-4	8.782e+5	0.149	32	0.858	1024	1	512	False
Bayesian Hidden+Output	256	2.710e-4	9.912e+5	0.149	32	0.858	1024	1	512	False
Bayesian RNN+Hidden +Output	512	1.488e-3	6.342e+5	0.252	32	1.291	16	1	0	True
Fully Bayesian	128	1.265e-3	9.983e+5	0.162	256	1.061	16	1	0	True