Adaptive Prediction Timing for Electronic Health Records

Jacob Deasy[*] **& Pietro Liò**
Department of Computer Science and Technology
University of Cambridge
Cambridge, United Kingdom
{jd645,pl219}@cam.ac.uk

Ari Ercole
Department of Medicine
University of Cambridge
Cambridge, United Kingdom
ae105@cam.ac.uk

Abstract

In realistic scenarios, multivariate timeseries evolve over case-by-case time-scales. This is particularly clear in medicine, where the rate of clinical events varies by ward, patient, and application. Increasingly complex models have been shown to effectively predict patient outcomes, but have failed to adapt granularity to these inherent temporal resolutions. As such, we introduce a novel, more realistic, approach to generating patient outcome predictions at an adaptive rate based on uncertainty accumulation in Bayesian recurrent models. We use a Recurrent Neural Network (RNN) and a Bayesian embedding layer with a new aggregation method to demonstrate adaptive prediction timing. Our model predicts more frequently when events are dense or the model is certain of event latent representations, and less frequently when readings are sparse or the model is uncertain. At 48 hours after patient admission, our model achieves equal performance compared to its static-windowed counterparts, while generating patient- and event-specific prediction timings that lead to improved predictive performance over the crucial first 12 hours of the patient stay.

1 Introduction

Over the past decade, machine learning, and deep learning in particular, have repeatedly demonstrated strong performance on a range of benchmark datasets (LeCun et al. (2015); Goodfellow et al. (2016)). These successes have since been transferred to medical data (Esteva et al. (2019)), and specifically the domain of Electronic Health Records (EHRs, Shickel et al. (2017)), where it is hoped research will lead to patient specification prognostication and diagnosis.

Recent state-of-the-art deep neural network (DNN) models for EHR data use recurrent models for sequence analysis which rely upon fixed prediction scheduling, carrying out extensive model analysis while overlooking the underlying choice of time window length (Choi et al. (2016); Rajkomar et al. (2018); Tomašev et al. (2019)). This early-stage modelling decision—necessitated by traditional Recurrent Neural Network (RNN, Elman (1990)) structure—loses patient information and timeseries granularity, and ignores the underlying timescales present in EHR data which have been shown to boost model performance (Meiring et al. (2018); Che et al. (2018)).

In this paper, we introduce and analyse a novel method for adaptive prediction timing in the context of medical timeseries. Progress in variational Bayesian neural networks, facilitated by the reparameterisation trick (Blundell et al. (2015); Kucukelbir et al. (2017)), has led to increased use of Bayesian embeddings of medical concepts (Dusenberry et al. (2019)). We posit that embedding distribution uncertainty can be used to induce adaptive prediction timing. To this end, we explore the following:

1. How can model uncertainty be related to prediction timing? (Section 2, paragraph 2)
2. How does adaptive prediction timing affect model performance? (Table 1)
3. How do prediction timings differ and adapt from fixed windows during training? (Figure 1)
4. Does our model generalise to different clinical objectives and cohorts? (Table 1 and A.2)

---

[*]corresponding author JD, https://github.com/jacobdeasy/dynamic-ehr

Contributions   We draw several significant conclusions from sequence modelling on the MIMIC-III (Johnson et al. (2016)) and the e-ICU (Pollard et al. (2018), Section A.2) datasets. We find that certainty rather than uncertainty, quantified by the precision of embedding distributions in a variational embedding layer, generates a natural measure of when to predict. In particular, we find that using the cumulative precision of embedding distributions encourages models to predict frequently when event sampling is dense and/or familiar to the model, and delays prediction when the model is uncertain about recent events—a more realistic approach to prediction timing. We demonstrate that the benefits of this model formulation do not impact negatively on final model performance. Finally, we highlight how adaptive prediction timing evolves over training to better utilise time periods of frequent events and produce correct predictions earlier in the patient stay.

2   BACKGROUND

Adaptive prediction timing   Recent, highly-cited, outcome prediction models for EHR timeseries (Rajkomar et al. (2018); Tomašev et al. (2019)) have paid little attention to prediction timing, relegating window choice to the supplementary material, and omitting the question '*When* is a good time to update model predictions?' from the line of enquiry. This approach contradicts evidence that modelling patient outcomes dynamically over time is beneficial (Meiring et al. (2018); Deasy et al. (2019)), and overlooks literature on adaptive computation for RNNs (Graves (2016)). A few efforts to overcome irregular sampling have been made by learning interpolants and decay rates for individual variables across fixed time periods (Che et al. (2018); Shukla & Marlin (2019)), but these models still do not account for varying amounts of *information* lost at the point of aggregation. The authors of Liu et al. (2019) recently demonstrated that dynamic prediction, with a patient-specific temporal resolution found by classical min-max optimisation, outperforms the previous one-size-fits-all approach despite maintaining fixed windows. In this paper, we go further by arguing that the patient timeseries is, in fact, an *information series* and it is more appropriate to evenly spread events based on model certainty. Our approach not only generates individualised prediction timing, but also *event-specific* prediction timing—crucial in the highly heterogeneous and patient-specific environment of the hospital and the real world.

Embedding precision   In a Bayesian RNN, the variational inference approach to learning the weights $\boldsymbol{w}$ of the approximate model $q(\boldsymbol{w}|\boldsymbol{\theta})$, dictates the use of factorised weight posteriors $q(\boldsymbol{w}|\boldsymbol{\theta}) = \prod_i q(\boldsymbol{w}_i|\boldsymbol{\theta}_i)$. When the $\boldsymbol{w}_i$ follow a multivariate Gaussian distribution, with learnable mean vector $\boldsymbol{\mu}$ and diagonal covariance matrix $\boldsymbol{\Sigma}$, we have

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Phi}^{-1}), \tag{1}$$

where we drop the index for ease of notation and note that $\boldsymbol{\Phi}$, the inverse of the covariance matrix $\boldsymbol{\Sigma}$, is the precision matrix of the multivariate normal. The precision of each embedding distribution is, therefore, defined by

$$\text{Precision}(\boldsymbol{w}) = \det(\boldsymbol{\Phi}) = \prod_j \boldsymbol{\Sigma}_{jj}^{-2}, \tag{2}$$

and is as a measure of model certainty (DeGroot (2005)).

3   ADAPTIVE PREDICTION TIMING

Clinical objectives   We analyse the performance of a novel recurrent model on in-hospital mortality and long length of stay (defined as greater than 7 days, Rajkomar et al. (2018)) prediction, both at 48 hours after admission. Dynamic mortality risk estimation helps summarise the patient state, predict patient trajectory, and is the subject of multiple clinical severity scores (Rapsang & Shyam (2014)). Equally, long length of stay estimation enables ward management planning and resource allocation across the hospital. We employ the MIMIC-III database (Johnson et al. (2016)), an EHR dataset collected from 46,520 patients admitted to intensive care units (ICUs) at Beth Israel Deaconess Medical Center. We embed all chart, lab, and output events as described in Deasy et al. (2019), utilise our adaptive prediction timing aggregation step, feed the output to a layer-normalised LSTM (Ba et al. (2016); Hochreiter & Schmidhuber (1997)), and perform an affine transformation before a sigmoid output activation.

Table 1: Mean (and standard deviation) of metrics for the adaptive prediction timing model on the binary mortality and long length of stay tasks—max MCC over 100 thresholds is reported. Our model displays strong predictive performance, with robust generalisation to the held-out test set.

TASK	METRIC	VALIDATION	TEST
Mortality	AUPRC	0.576 ($\pm$0.013)	0.556 ($\pm$0.011)
	AUROC	0.897 ($\pm$0.003)	0.879 ($\pm$0.004)
	MCC	0.510 ($\pm$0.011)	0.496 ($\pm$0.012)
Long length of stay	AUPRC	0.614 ($\pm$0.008)	0.566 ($\pm$0.009)
	AUROC	0.834 ($\pm$0.004)	0.830 ($\pm$0.004)
	MCC	0.494 ($\pm$0.009)	0.465 ($\pm$0.009)

Method For a given patient sequence, comprised of time points and events $\{t_i, x_i\}_{0 \leq i \leq n}$, instead of aggregating events into fixed time intervals, we first sample the corresponding sequence of event embedding distributions to obtain the sequence $\{\boldsymbol{w}_i\}_{0 \leq i \leq n}$. To separate these samples into intervals, we then use the embedding distributions to generate a corresponding cumulative precision sequence

$$\boldsymbol{p}_k^* = \sum_{i=0}^{k} \boldsymbol{p_i} = \sum_{i=0}^{k} \left(\prod_j (\boldsymbol{\Sigma_i})_{jj}^{-2} \right), \tag{3}$$

and separate this sequence into *equi-precise* aggregation windows which evolve as embedding distributions are refined by training. At no point does our model make use of event timestamps. We implement this in a vectorised manner to handle batches of size greater than one.

Our models were trained by minimising the Kullback-Leibler (KL) divergence between the approximate posterior and the actual, intractable, posterior via the reparameterisation trick. Equivalent to minimising an expectation over the negative log-likelihood term plus a KL regularisation term

$$\mathcal{L}(\boldsymbol{\theta}) = \text{KL}[q(\boldsymbol{w}|\boldsymbol{\theta}) \parallel p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X})] \tag{4}$$

$$\propto \text{KL}[q(\boldsymbol{w}|\boldsymbol{\theta}) \parallel p(\boldsymbol{w})] - \mathbb{E}_q[\ln p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})]. \tag{5}$$
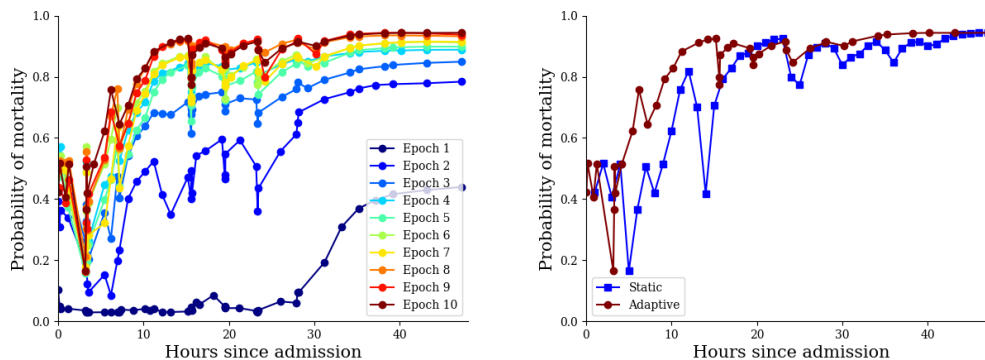
4 EXPERIMENTS

Clinical tasks For our clinical tasks, to assess predictive performance, we measure area under the precision-recall curve (AUPRC), area under the receiver operating characteristic curve (AUROC), and, as there is a strong class imbalance (see Table A.1), Matthews Correlation Coefficient (MCC). Table 1 shows the mean and standard deviation of the metrics at 48 hours after admission. Throughout, we re-sample the embedding layer of the variational models 100 times and bootstrapped ensembles of 10 deterministic models with 1000 re-samples to generate error measurements. Despite the underlying change in aggregation mechanism, the final performance of our model is very strong, inline with the literature (Dusenberry et al. (2019)), and the predictions are well-calibrated (see Figure A.1), demonstrating strong performance on both clinical tasks. We also verify model generalisation for both tasks on the eICU dataset (Pollard et al. (2018)) in Table A.2.

Model variants We assess our model against a range of differing models in Table 2. We name these models based on whether their embedding layer is deterministic or Bayesian, and whether their prediction timings are based on fixed timing (no prefix), fixed event count (#), or fixed cumulative precision (p^*). We include a model which aggregates by event count, to isolate the effect of our model's adaptation, and conclude it has a marginally negative effect on generalisation, which must be weighed against the advantages of improved temporal resolution during eventful periods.

Early predictive power In Figure 1, we compare mortality risk prediction of the static model with our adaptive model for a patient who went on to die in hospital. The more fine-grained prediction timings learnt by the variational model, displayed in Figure 1-left, led to earlier prediction of mortality compared to the static model in Figure 1-right due to patient-specific segmentation of the timeseries. As most patients in the ICU have many additional readings taken in the first hours of

Table 2: Performance comparison between different model variants for the binary mortality prediction tasks on the MIMIC-III dataset.

MODEL	VAL. AUPRC	VAL AUROC	TEST AUPRC	TEST AUROC
Deterministic LSTM	0.592 ($\pm$0.010)	0.887 ($\pm$0.008)	0.595 ($\pm$0.012)	0.889 ($\pm$0.006)
Deterministic #-LSTM	0.602 ($\pm$0.012)	0.883 ($\pm$0.006)	0.578 ($\pm$0.011)	0.883 ($\pm$0.005)
Bayesian LSTM	0.574 ($\pm$0.013)	0.886 ($\pm$0.004)	0.571 ($\pm$0.012)	0.883 ($\pm$0.004)
Bayesian #-LSTM	0.582 ($\pm$0.011)	0.889 ($\pm$0.004)	0.573 ($\pm$0.013)	0.881 ($\pm$0.004)
Bayesian p^*-LSTM	0.576 ($\pm$0.013)	0.897 ($\pm$0.003)	0.556 ($\pm$0.011)	0.879 ($\pm$0.004)

Figure 1: **Left**–Mortality probability evolving through training for a patient that went on to die. **Right**–The adaptive timing model predicts mortality earlier than the static window model.

their stay (e.g. admission information and medical history), which clinicians use to more frequently update their opinion of the patient state, this is a more realistic approach to outcome prediction.

Prediction timing evolution In Figure 2, we demonstrate the evolution of prediction timing for a different patient. In Figure 2-right, the prediction timing distribution can be seen to focus on a particularly event-dense period for this patient as it learns to be more certain about the embedding distributions of particular clinical events. This suggests our model would adapt well to more extreme shifts in granularity such as stays which include either surgical or emergency interventions.
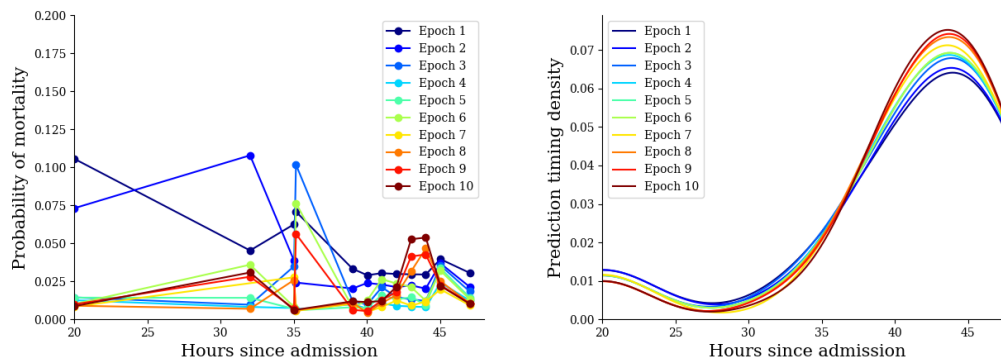

Figure 2: Example of how the timing of model predictions coalesce through training for a patient with frequent event recordings towards the end of their stay.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pp. 3504–3512, 2016.

Jacob Deasy, Pietro Liò, and Ari Ercole. Dynamic survival prediction in intensive care units from heterogeneous time series without the need for variable selection or pre-processing. *arXiv preprint arXiv:1909.07214*, 2019.

Morris H DeGroot. *Optimal statistical decisions*, volume 82. John Wiley & Sons, 2005.

Michael W Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M Dai. Analyzing the role of model uncertainty for electronic health records. *arXiv preprint arXiv:1906.03842*, 2019.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24, 2019.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, 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, 2016.

Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures. In *International conference on machine learning*, pp. 2342–2350, 2015.

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

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.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

Luchen Liu, Haoran Li, Zhiting Hu, Haoran Shi, Zichang Wang, Jian Tang, and Ming Zhang. Learning hierarchical representations of electronic health records for clinical outcome prediction. *arXiv preprint arXiv:1903.08652*, 2019.

Christopher Meiring, Abhishek Dixit, Steve Harris, Niall S MacCallum, David A Brealey, Peter J Watkinson, Andrew Jones, Simon Ashworth, Richard Beale, Stephen J Brett, et al. Optimal intensive care outcome prediction over time using machine learning. *PloS one*, 13(11):e0206862, 2018.

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.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

Amy Grace Rapsang and Devajit C Shyam. Scoring systems in the intensive care unit: a compendium. *Indian journal of critical care medicine: peer-reviewed, official publication of Indian Society of Critical Care Medicine*, 18(4):220, 2014.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

Satya Narayan Shukla and Benjamin M Marlin. Interpolation-prediction networks for irregularly sampled time series. *arXiv preprint arXiv:1909.07782*, 2019.

Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

A APPENDIX

A.1 ADDITIONAL DATASET INFORMATION

After applying the pipeline described in Deasy et al. (2019), our subset of the MIMIC-III dataset contained 21,143 patient stays including the following demographic and outcome ratios.

Table A.1: Dataset information.

	TRAIN (%)	VALIDATION (%)	TEST (%)
Male	9329/16913 (55.2%)	1140/2115 (53.9%)	1162/2115 (54.9%)
In-hospital mortality	2237/16913 (13.2%)	280/2115 (13.2%)	280/2115 (13.2%)
Long length of stay	3698/16913 (21.9%)	512/2115 (24.2%)	462/2115 (21.8%)

A.2 EICU COLLABORATIVE RESEARCH DATABASE

As a proof of concept that our findings generalise, we also experiment with a small subset of the aperiodic vital sign readings in the eICU Collaborative Research Database (eICU) dataset (Pollard et al. (2018)), another publicly available EHR dataset. Results are displayed in Table A.2.

Table A.2: Performance of the adaptive prediction timing model for the binary mortality and long length of stay tasks on the eICU dataset.

TASK	METRIC	VALIDATION	TEST
Mortality	AUPRC	0.253 ($\pm$ 0.012)	0.244 ($\pm$ 0.010)
	AUROC	0.705 ($\pm$ 0.005)	0.701 ($\pm$ 0.006)
Long length of stay	AUPRC	0.217 ($\pm$ 0.010)	0.205 ($\pm$ 0.012)
	AUROC	0.610 ($\pm$ 0.005)	0.600 ($\pm$ 0.007)

A.3 MODEL CALIBRATION

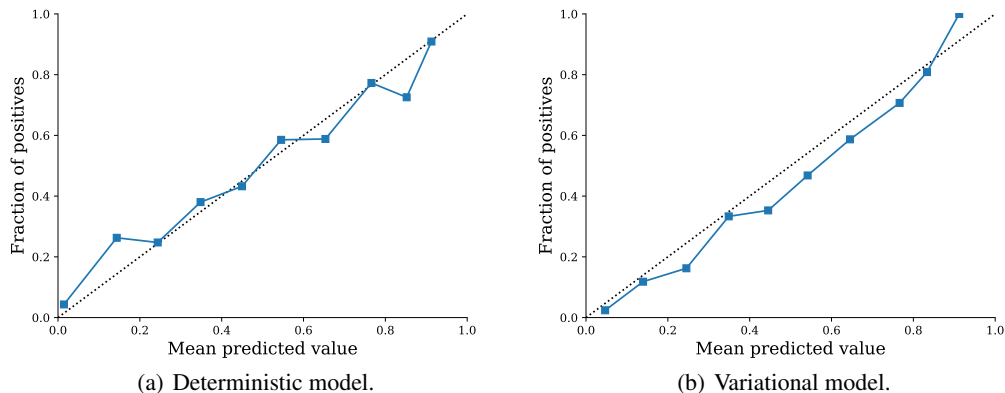(a) Deterministic model.

(b) Variational model.

Figure A.1: Model calibration curves.

A.4 ADDITIONAL TRAINING DETAILS

As in Deasy et al. (2019), our model uses embeddings of raw, unprocessed medical concepts in order to avoid any loss of extraneous information. Continuous variables were binned into 10 discrete categories assigned by quantile, and discrete variables were left untouched. Missing events were also embedded as discrete categories so our model makes use of *informative missingness* (Che et al. (2018)). Following on from the embedding layer, we noted during experimentation that a layer-normalised LSTM variant led to a considerable increase in performance. We initialise our LSTM with glorot initialisation (Glorot & Bengio (2010)) for the input-to-hidden matrices, orthogonal initialisation for the hidden-to-hidden matrices, and set the forget gate bias to 1 before training (Jozefowicz et al. (2015)).

For each model, we search over the space of hyperparameters defined in Table A.3 using the Python package `wandb` from Weights and Biases. The search was performed using Bayesian optimisation to minimise validation set loss over both discrete and continuous variables. We use HyperBand early stopping (Li et al. (2017)), with $s = 2$, $\eta = 3$, and $max\_iter = 10$ to expedite optimisation. We fix batch size at 64 for all models.

Models were implemented in PyTorch 1.4, and trained on a Nvidia Titan X using Adam optimisation (Kingma & Ba (2014)). Both e-ICU and MIMIC-III datasets were split into train, validation, and test sets in a ratio of 8:1:1.

Table A.3: Hyperparameter ranges and options.

HYPERPARAMETER	RANGE
Learning rate	[0.00001, 0.1]
L_2 regularisation coefficient	[0.0, 0.01]
Prior standard deviation (Bayesian only)	[0.1, 1.0]
Embedding dimension	[16, 32, 48, 64]
LSTM hidden dimension	[16, 32, 64, 128, 256, 512]