Multimedia Appendix 2: Experimental settings

Experimental settings
For our experiments, we split our training dataset (1642 examples) into 75%, 15% and 10% to form our train (1232 examples), validation (246 examples) and internal test dataset (164 examples). Hyperparameters are tuned on the validation dataset.

Experimental settings for multi-task architecture
Our implementation of Multi-task architecture is based on [1][2]. We trained our model on two NVIDIA(R) V100 GPU using the PyTorch framework. As we are using BERT-base architecture, all the texts were tokenized using WordPieces[3] and tokenized text were chopped to spans no longer than 512 tokens. We used Adamax [4] as our optimizer with a learning rate of 5e-5 and a batch size of 32 by following [5]. The maximum number of epochs (epochmax) was set to 100. A linear learning rate decay schedule with warm-up over 0.1 was used. To avoid the exploding gradient problem, we clipped the gradient norm within 1. We use the same hyperparameters for all the task heads. The hyperparameters are summarized in Table 1.

| Hyperparameter | Value  |
|----------------|--------|
| Learning Rate  | 5e-5   |
| Batch Size     | 32     |
| epoch\textsuperscript{max} | 100     |
| Dropout        | 0.1    |
| Optimizer      | Adamax |

Table 1. Hyperparameters for multi-task learning.

Experimental settings for fine-tuning
To fine-tune the IIT-MTL-ClinicalBERT on specific tasks, we change the maximum number of epochs to 10 and learning rate to 1e−5. All the hyperparameters are summarized in Table 2.

| Hyperparameter | Value  |
|----------------|--------|
| Learning Rate  | 1e-5   |
| Batch Size     | 32     |
| epoch\textsuperscript{max} | 10     |
| Dropout        | 0.1    |
| Optimizer      | Adamax |

Table 2. Hyperparameters for Fine-tuning.

Experimental settings for the ensemble module
Table 3 provides the parameters used for Bayesian regression and ridge regression.

| Hyperparameter | Value  |
|----------------|--------|
|               |        |

Table 3. Experimental settings for Bayesian regression and ridge regression.
| Technique         | Parameter           | Value       |
|-------------------|---------------------|-------------|
| Bayesian Regression| Number of iterations| 300         |
|                   | alpha_1<sup>a</sup> | 1.15e-06    |
|                   | alpha_2<sup>b</sup> | 1.02e-06    |
|                   | lambda_1<sup>c</sup>| 1e-06       |
|                   | rate<sup>d</sup>    | 1e-06       |
| Ridge Regression  | alpha               | 1           |
|                   | solver              | auto<sup>e</sup> |

<sup>a</sup> shape parameter for the Gamma distribution prior over the alpha parameter  
<sup>b</sup> rate parameter for the Gamma distribution prior over the alpha parameter  
<sup>c</sup> shape parameter for the Gamma distribution prior over the lambda parameter  
<sup>d</sup> parameter for the Gamma distribution prior over the lambda parameter  
<sup>e</sup> chooses the solver automatically based on the type of data

**References**

1. Liu X, He P, Chen W, Gao J. Multi-task deep neural networks for natural language understanding. ACL 2019 - 57th Annu Meet Assoc Comput Linguist Proc Conf 2019;4487–4496.

2. namisan/mt-dnn: Multi-Task Deep Neural Networks for Natural Language Understanding [Internet]. [cited 2020 Nov 2]. Available from: https://github.com/namisan/mt-dnn

3. [61] Wu Y, Schuster M, Chen Z, Le Q V., Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser Ł, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J. Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. 2016;1–23. Available from: http://arxiv.org/abs/1609.08144

4. Kingma DP, Ba JL. Adam: A method for stochastic optimization. 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc 2015.

5. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol - Proc Conf 2019;1(Mlm):4171–4186.