SUPPLEMENT TO 'INDIVIDUALIZED RISK ASSESSMENT OF PREOPERATIVE OPIOID USE BY INTERPRETABLE NEURAL NETWORK REGRESSION'

BY YUMING SUN[1], JIAN KANG[1,a], CHAD BRUMMETT[2] AND YI LI[1,b]

[1]*Department of Biostatistics, University of Michigan, Ann Arbor;* [a]*jiankang@umich.edu;* [b]*yili@umich.edu*

[2]*Department of Anesthesiology, University of Michigan, Ann Arbor*

Section A: Architecture of DNN

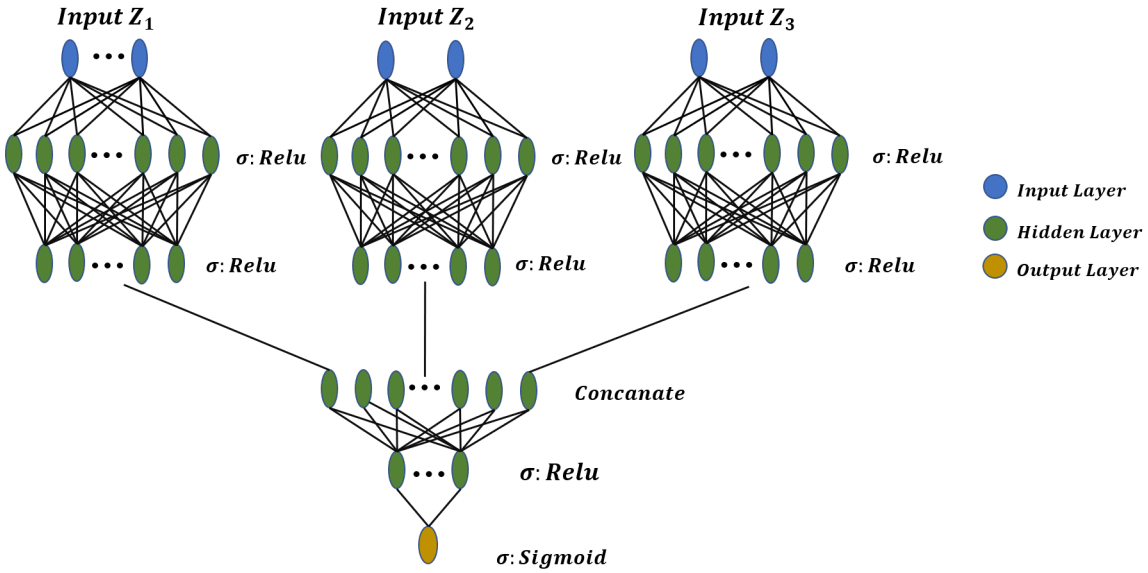

Fig 1: **Architecture of DNN for the AOS Data. Input:** preoperative characteristics are classified into three inputs: Z_1 are un-modifiable characteristics, such as gender and race; Z_2 are modifiable characteristics, such as BMI and smoking; Z_3 are pain-related characteristics, such as Fibromyalgia Survey Score and pain severity. **Layers:** each category of inputs goes through the same structure: two hidden layers with a ReLu activation function. The first hidden layer has 500 neurons and the the second hidden layer has 125 neurons. The three structures are concatenated and passed onto a layer with 15 hidden neurons and a ReLu activation function. **Output:** estimated probability of preoperative opioid use.

Section B: Sensitivity Analysis

Because stochastic gradient descent is sensitive to the choice of learning rates (LR), we use grid search to tune the learning rate. For the real data analysis, we tune the learning rate over a range from 0.005 to 0.1 with 20 equally spaced grid points, and find that $LR = 0.01$ seems to strike a balance between stability and computational readiness. We also implement the adaptive SGD to analyze our data. Specifically, we have implemented three popular adaptive

Keywords and phrases: deep learning, pain research, precision medicine, generalized linear models.

SGD algorithms, namely, "Adagrad", "Adadelta" and "Adam." Adagrad adapts the learning rate based on a sequence of subgradients [Duchi, Hazan and Singer (2011)] to improve the robustness of SGD and avoid tuning the learning rate manually [Dean et al. (2012)], while Adadelta [Zeiler (2012)] and Adam [Kingma and Ba (2014)] only store an exponentially decaying average of subgradients [Zeiler (2012)]. We conduct 100 experiments to compare the prediction performance using different optimizers. In each experiment, we randomly split data into the training and testing parts, and use the balanced subsampling strategy described in Section 5.2 to assess the performance on the testing data. The means and standard errors (se) of different metrics are summarized in Table 1. We find that all four methods give similar performances, though SGD with a fixed $LR = 0.01$ and Adam give the same C-statistic and sensitivity, slightly better than those obtained by Adagrad and Adadelta; all of these methods give the same balance accuracy.

TABLE 1
Prediction Performance of INNER Using different Optimizers[a,b]

	SGD (LR=0.01)	**Adagrad**	**Adadelta**	**Adam**
C-statistic	0.78 (0.0006)	0.77 (0.0005)	0.76 (0.0006)	0.78 (0.0006)
Accuracy	0.72 (0.0029)	0.73 (0.0011)	0.73 (0.0006)	0.72 (0.0009)
Sensitivity	0.69 (0.0052)	0.66 (0.0022)	0.66 (0.0012)	0.69 (0.0020)
Specificity	0.73 (0.0052)	0.76 (0.0019)	0.76 (0.0010)	0.73 (0.0017)
Balance Accuracy	0.71 (0.0008)	0.71 (0.0006)	0.71 (0.0005)	0.71 (0.0006)

a. used the balanced subsampling strategy and a threshold of 0.5
b. based on 100 random splits

The number of iterations is chosen to ensure the convergence of the algorithm (as shown in Fig 2). We have also varied the batch sizes and number of iterations to examine the stability of the results and find a batch size of 64 and an epoch of 200 give a reasonable performance. We have conducted sensitivity analysis to assess the robustness of SGD towards the choices of these hyperparameters, and we find that the model's C-statistic is fairly robust to them. Specifically, by varying the learning rate from 0.0075 to 0.0125, the batch size from 32 to 128 and the number of iterations from 200 to 250, the C-statistic of the obtained INNER model is around 0.78.

TABLE 2
Average C-statistics (se) of INNER with Various Learning Rates, Batch Sizes and Epochs[a,b,c]

		LR =0.0075	**LR = 0.01**	**LR = 0.0125**
BS = 32	Epoch = 150	0.78 (0.0005)	0.78 (0.0006)	0.78 (0.0006)
	Epoch = 200	0.78 (0.0005)	0.78 (0.0007)	0.78 (0.0006)
	Epoch = 250	0.78 (0.0005)	0.78 (0.0007)	0.78 (0.0006)
BS = 64	Epoch = 150	0.78 (0.0006)	0.78 (0.0007)	0.78 (0.0007)
	Epoch = 200	0.78 (0.0005)	0.78 (0.0006)	0.78 (0.0006)
	Epoch = 250	0.78 (0.0005)	0.78 (0.0006)	0.78 (0.0005)
BS =128	Epoch = 150	0.78 (0.0006)	0.78 (0.0006)	0.78 (0.0006)
	Epoch = 200	0.78 (0.0006)	0.78 (0.0006)	0.78 (0.0006)
	Epoch = 250	0.78 (0.0006)	0.78 (0.0005)	0.78 (0.0006)

a. used the balanced subsampling strategy and a threshold of 0.5
b. used SGD for optimization
c. based on 100 experiments

We have conducted additional sensitivity analyses to examine the performance of the model under various initialization schemes for the weights $\mathbf{W}$ and the biases $\mathbf{b}$ in the neural

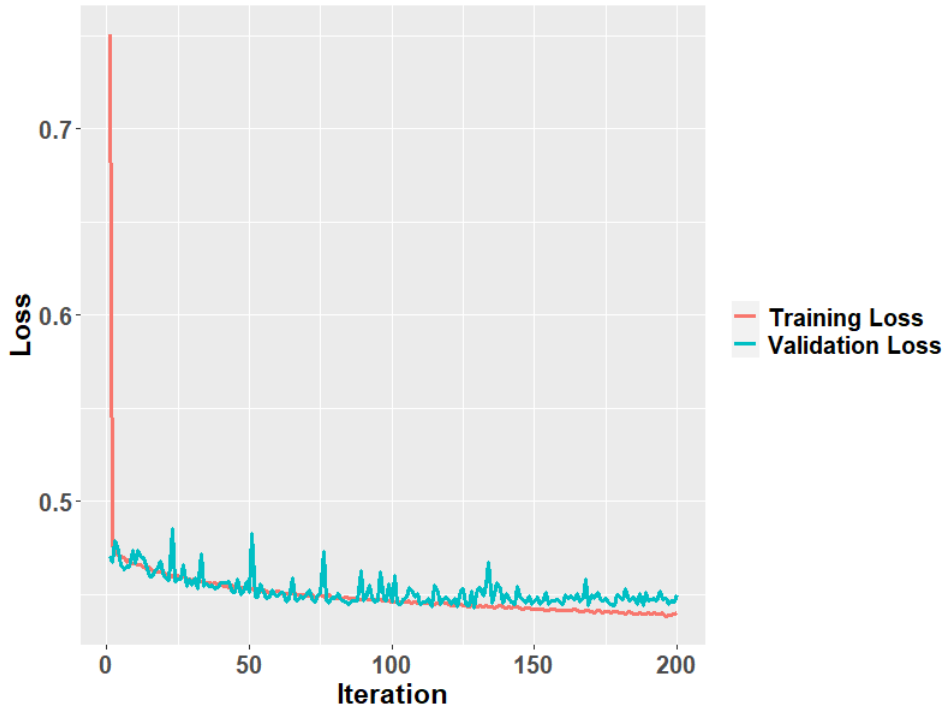

Fig 2: **Learning Curve of INNER: Cross Entropy Loss Against Iteration For Training and Validation Data**

networks. We have explored using different weights, such as uniform and normal weights, for the initial weights [Glorot and Bengio (2010); He et al. (2015)]. In particular, we have studied two versions of uniform weights: for a weight matrix $\mathbf{W}_l \in \mathbf{R}^{k_{l+1} \times k_l}$, where k_l and k_{l+1} are the numbers of input and output units of the lth layer, we initialize it with Uniform$\{-\sqrt{6/(k_l + k_{l+1})}, \sqrt{6/(k_l + k_{l+1})}\}$ following Glorot and Bengio (2010) (labeled as "Glorot uniform" in Table 3, which reports the sensitivity analysis results); we also initialize the weight matrix with Uniform$(-\sqrt{6/k_l}, \sqrt{6/k_l})$ following He et al. (2015) (labeled as "He uniform" in Table 3). For the normal weights, we use Normal$(0, 2/(k_l + k_{l+1}))$ as the initial weights following Glorot and Bengio (2010) (labeled as "Glorot normal" in Table 3). Finally, for the bias vector $\mathbf{b}$, we initialize it to be either all 0's or 1's for its components (labeled as "Zeros" or "Ones" in the column of bias initialization in Table 3). For each set-up, we find that the C-statistic of the model is fairly constant, which is 0.78 with varied initialized values of weights and biases.

TABLE 3
Average (se) C-statistics with different Initializations of Weights and Biases[a,b,c]

Weight Initialization	Bias Initialization	**C-statistic**
Glorot uniform	Zeros	0.78 (0.0006)
	Ones	0.78 (0.0006)
Glorot normal	Zeros	0.78 (0.0005)
	Ones	0.78 (0.0005)
He uniform	Zeros	0.78 (0.0008)
	Ones	0.78 (0.0006)

a. used the balanced subsampling strategy and a threshold of 0.5
b. used SGD for optimization
c. based on 100 experiments

TABLE 4
Comparisons of the Prediction Performance using the AOS Data[a,b,c]

	Deep Neural Network	Logistic Regression	Interpretable Neural Network Regression
Preoperative Opioid Prevalence: 0.23			
C-statistic	0.78 (0.0006)	0.62 (0.0094)	0.78 (0.0006)
Threshold = 0.50			
Accuracy	0.80 (0.0004)	0.70 (0.0116)	0.80 (0.0004)
Sensitivity	0.33 (0.0049)	0.43 (0.0331)	0.31 (0.0057)
Specificity	0.94 (0.0016)	0.78 (0.0238)	0.94 (0.0018)
Balance Accuracy	0.63 (0.0017)	0.61 (0.0071)	0.63 (0.002)
Threshold = 0.23			
Accuracy	0.72 (0.0021)	0.69 (0.0123)	0.73 (0.0030)
Sensitivity	0.69 (0.0039)	0.44 (0.0336)	0.68 (0.0055)
Specificity	0.73 (0.0038)	0.77 (0.025)	0.74 (0.0054)
Balance Accuracy	0.71 (0.0006)	0.61 (0.007)	0.71 (0.0007)
Preoperative Opioid Prevalence: 0.50			
C-statistic	0.78 (0.0006)	0.73 (0.0027)	0.78 (0.0006)
Threshold = 0.50			
Accuracy	0.73 (0.0017)	0.63 (0.0129)	0.72 (0.0029)
Sensitivity	0.69 (0.0043)	0.67 (0.0261)	0.69 (0.0052)
Specificity	0.73 (0.0034)	0.62 (0.0238)	0.73 (0.0052)
Balance Accuracy	0.71 (0.0007)	0.64 (0.0049)	0.71 (0.0008)
Threshold = 0.23			
Accuracy	0.46 (0.0044)	0.50 (0.0130)	0.41 (0.0056)
Sensitivity	0.93 (0.0024)	0.84 (0.0154)	0.95 (0.0022)
Specificity	0.31 (0.0064)	0.39 (0.0211)	0.24 (0.0080)
Balance Accuracy	0.62 (0.0021)	0.61 (0.0047)	0.60 (0.0030)

a. prediction power of each model with the best architectures (DNN and INNER) under different sampling strategies and threshold; for the comparison of different architectures, refer to Appendix Table 5 and Appendix Table 6

b. based on 100 experiments for each metric

c. in the AOS data, the prevalence of preoperative opioid is 0.23, and the prevalence is around 0.23 for the training data; we use the balanced subsampling strategy to adjust the prevalence of preoperative opioid to be 0.50 in the training data

TABLE 5
Tuning the Architecture of INNER with the AOS Data[a,b,c,d]

	Three Layers 250 Neurons	**Four Layers 500 Neurons**	**Five Layers 500 Neurons**
Preoperative Opioid Prevalence: 0.23			
C-statistic	0.78 (0.0006)	0.78 (0.0007)	0.77 (0.0005)
Threshold = 0.50			
Accuracy	0.80 (0.0004)	0.79 (0.0005)	0.79 (0.0004)
Sensitivity	0.31 (0.0057)	0.32 (0.0063)	0.32 (0.0061)
Specificity	0.94 (0.0018)	0.94 (0.0021)	0.93 (0.0019)
Balance Accuracy	0.63 (0.0020)	0.63 (0.0021)	0.63 (0.0021)
Threshold = 0.23			
Accuracy	0.73 (0.0030)	0.72 (0.0031)	0.72 (0.0022)
Sensitivity	0.68 (0.0055)	0.68 (0.0054)	0.68 (0.0043)
Specificity	0.74 (0.0054)	0.73 (0.0055)	0.74 (0.0041)
Balance Accuracy	0.71 (0.0007)	0.71 (0.0008)	0.71 (0.0005)
Preoperative Opioid Prevalence: 0.50			
C-statistic	0.78 (0.0006)	0.78 (0.0006)	0.78 (0.0006)
Threshold = 0.50			
Accuracy	0.72 (0.0029)	0.73 (0.0026)	0.72 (0.0020)
Sensitivity	0.69 (0.0052)	0.69 (0.0048)	0.69 (0.0037)
Specificity	0.73 (0.0052)	0.75 (0.0047)	0.73 (0.0036)
Balance Accuracy	0.71 (0.0008)	0.71 (0.0008)	0.71 (0.0005)
Threshold = 0.23			
Accuracy	0.41 (0.0056)	0.42 (0.0057)	0.43 (0.0050)
Sensitivity	0.95 (0.0022)	0.95 (0.0023)	0.94 (0.0021)
Specificity	0.24 (0.0080)	0.26 (0.0081)	0.28 (0.0071)
Balance Accuracy	0.60 (0.0030)	0.60 (0.0030)	0.61 (0.0026)

a. the first column is for the best INNER architecture as reported in Table 4 in the main text and Table 4

b. the other columns refer to the other more complicated INNERs, with more hidden layers or more neurons in each hidden layers

c. the column names are the number of hidden layers and the number of neurons in the first hidden layers for $F_L(\mathbf{Z}_i;\boldsymbol{\alpha})$ and $F_L(\mathbf{Z}_i;\boldsymbol{\alpha})$

d. in the AOS data, the prevalence of preoperative opioid use is 0.23; uses a balanced subsampling strategy by over-sampling cases; adjusts the prevalence of preoperative opioid use to be 0.50 in the training data

TABLE 6
Tuning the Architecture of DNN for AOS Data[a,b,c,d]

	Two Layers 500 Neurons	Three Layer 250 Neurons	Three Layer 500 Neurons
Preoperative Opioid Prevalence: 0.23			
C-statistic	0.78 (0.0006)	0.79 (0.0006)	0.79 (0.0005)
Threshold = 0.50			
Accuracy	0.80 (0.0004)	0.79 (0.0004)	0.79 (0.0004)
Sensitivity	0.33 (0.0049)	0.34 (0.0054)	0.32 (0.0068)
Specificity	0.94 (0.0016)	0.93 (0.0018)	0.94 (0.0020)
Balance Accuracy	0.63 (0.0017)	0.63 (0.0018)	0.63 (0.0024)
Threshold = 0.23			
Accuracy	0.72 (0.0021)	0.72 (0.0022)	0.72 (0.0024)
Sensitivity	0.69 (0.0039)	0.70 (0.0038)	0.69 (0.0049)
Specificity	0.73 (0.0038)	0.72 (0.0040)	0.73 (0.0046)
Balance Accuracy	0.71 (0.0006)	0.71 (0.0006)	0.71 (0.0005)
Preoperative Opioid Prevalence: 0.50			
C-statistic	0.78 (0.0006)	0.78 (0.0006)	0.78 (0.0005)
Threshold = 0.50			
Accuracy	0.73 (0.0017)	0.72 (0.0025)	0.72 (0.0023)
Sensitivity	0.69 (0.0043)	0.70 (0.0039)	0.70 (0.0043)
Specificity	0.73 (0.0034)	0.72 (0.0043)	0.73 (0.0042)
Balance Accuracy	0.71 (0.0007)	0.71 (0.0007)	0.71 (0.0005)
Threshold = 0.23			
Accuracy	0.46 (0.0044)	0.45 (0.0044)	0.45 (0.0052)
Sensitivity	0.93 (0.0024)	0.94 (0.0020)	0.93 (0.0023)
Specificity	0.31 (0.0064)	0.30 (0.0062)	0.31 (0.0074)
Balance Accuracy	0.62 (0.0021)	0.62 (0.0022)	0.62 (0.0026)

a. the first column is for the best DNN architecture as reported in Table 4 in the main text and Appendix Table 4

b. the other columns refer to the other more complicated DNNs, with more hidden layers or more neurons

c. the column names are the number of hidden layers and the number of neurons in the first hidden layers before concatenation

d. in the AOS data, the prevalence of preoperative opioid use is 0.23; uses a balanced subsampling strategy by over-sampling cases; adjusts the prevalence of preoperative opioid use to be 0.50 in the training data

REFERENCES

DEAN, J., CORRADO, G., MONGA, R., CHEN, K., DEVIN, M., MAO, M., RANZATO, M., SENIOR, A., TUCKER, P., YANG, K. and NG, A. Y. (2012). Large scale distributed deep networks. *Advances in Neural Information Processing Systems* **25** 1223–1231.

DUCHI, J., HAZAN, E. and SINGER, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* **12** 2121–2159.

GLOROT, X. and BENGIO, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 249–256. JMLR Workshop and Conference Proceedings.

HE, K., ZHANG, X., REN, S. and SUN, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision* 1026–1034.

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

ZEILER, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.