**SUPPLEMENTARY**

Table 1. Qualifying ICD-9 codes for heart failure

|  |  |
| --- | --- |
| **ICD-9 Code** | **Description** |
| 398.91 | Rheumatic heart failure (congestive) |
| 402.01 | Malignant hypertensive heart disease with heart failure |
| 402.11 | Benign hypertensive heart disease with heart failure |
| 402.91 | Unspecified hypertensive heart disease with heart failure |
| 404.01 | Hypertensive heart and chronic kidney disease, malignant, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified |
| 404.03 | Hypertensive heart and chronic kidney disease, malignant, with heart failure and with chronic kidney disease stage V or end stage renal disease |
| 404.11 | Hypertensive heart and chronic kidney disease, benign, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified |
| 404.13 | Hypertensive heart and chronic kidney disease, benign, with heart failure and chronic kidney disease stage V or end stage renal disease |
| 404.91 | Hypertensive heart and chronic kidney disease, unspecified, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified |
| 404.93 | Hypertensive heart and chronic kidney disease, unspecified, with heart failure and chronic kidney disease stage V or end stage renal disease |
| 428.0 | Congestive heart failure, unspecified |
| 428.1 | Left heart failure |
| 428.20 | Systolic heart failure, unspecified |
| 428.21 | Acute systolic heart failure |
| 428.22 | Chronic systolic heart failure |
| 428.23 | Acute on chronic systolic heart failure |
| 428.30 | Diastolic heart failure, unspecified |
| 428.31 | Acute diastolic heart failure |
| 428.32 | Chronic diastolic heart failure |
| 428.33 | Acute on chronic diastolic heart failure |
| 428.40 | Combined systolic and diastolic heart failure, unspecified |
| 428.41 | Acute combined systolic and diastolic heart failure |
| 428.42 | Chronic combined systolic and diastolic heart failure |
| 428.43 | Acute on chronic combined systolic and diastolic heart failure |
| 428.9 | Heart failure, unspecified |

**Details of case-control matching**

Control subjects were required to have their first office encounter within one year of the matching HF case patient’s first office visit, and have at least one office encounter 30 days before or any time after the case’s HF diagnosis date to ensure similar duration of observations among cases and controls.

**6-fold Cross Validation Scheme**



Figure 9. Diagram of 6-fold cross validation

Figure 6 depicts the 6-fold cross validation we performed for HF prediction. At each fold, we take five chunks as the training set, one chunk as the validation set, and the remaining one chunk as the test set. The training set is used to train the models. We train the models for a fixed number of epochs, where a single epoch is a complete iteration over the training set. After each epoch, we evaluate the model’s AUC using the validation set. After the fixed number of epochs, we pick the model that showed the highest validation AUC and evaluate its AUC on the test set. We do this for 6 different folds and calculate the average test AUC, which is used as the final model performance.

**Hyper-parameters used for training models**

After experimenting with various values, the following hyper-parameter setting produced the best performance. We used Theano 0.7 and CUDA 7 for training logistic regression, MLP, and GRU models. SVM was implemented with Scikit-Learn Linear SVC. KNN was implemented with Scikit-Learn KNeighborsClassifier.

Table 2. Hyper-parameter settings for training the models

|  |  |
| --- | --- |
| Model | Hyper-parameter |
| Logistic regression,  one-hot vectors | L2 regularization: 0.1, Max epoch: 100 |
| Logistic regression,  grouped code vectors &  medical concept vectors | L2 regularization: 0.01, Max epoch: 100 |
| SVM,  one-hot vectors | L2 regularization: 0.000001, Dual: False |
| SVM,  grouped code vectors &  medical concept vectors | L2 regularization: 0.001, Dual: False |
| MLP,  one-hot vectors | L2 regularization: 0.01, Hidden layer size: 15, Max epoch: 100 |
| MLP,  grouped code vectors &  medical concept vectors | L2 regularization: 0.001, Hidden layer size: 100, Max epoch: 100 |
| KNN,  one-hot vectors | Number of neighbors: 15 |
| KNN,  grouped code vectors &  medical concept vectors | Number of neighbors: 100 |
| All GRU models | L2 regularization: 0.001, Hidden layer size: 100, Max epoch: 100 |

**AUC, standard error values of Figure 4 and Figure 5**

Table 3. AUC, standard error values of Figure 4.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Logistic Regression | SVM | MLP | KNN | GRU w/o duration | GRU w/ duration |
| One-hot encoding | 0.7096±0.0131 | 0.7105±0.0078 | 0.7184±0.0117 | 0.6118±0.0129 | 0.7493±0.0158 | 0.7540±0.0137 |
| Grouped code vectors | 0.7351±0.0200 | 0.7361±0.0187 | 0.7381±0.0182 | 0.6887±0.0125 | 0.7593±0.0163 | 0.7620±0.0164 |
| Medical concept vectors | 0.7467±0.0120 | 0.7435±0.0111 | 0.7649±0.0137 | 0.7293±0.0113 | 0.7701±0.0111 | 0.7768±0.0117 |

Table 4. AUC, standard error values of Figure 5.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Logistic Regression | SVM | MLP | KNN | GRU w/o duration | GRU w/ duration |
| One-hot encoding | 0.7896±0.0144 | 0.7826±0.0106 | 0.8029±0.0165 | 0.6779±0.0099 | 0.8556±0.0141 | 0.8633±0.0128 |
| Grouped code vectors | 0.8071±0.0147 | 0.8098±0.0165 | 0.8192±0.0101 | 0.7550±0.0116 | 0.8618±0.0079 | 0.8699±0.0086 |
| Medical concept vectors | 0.8092±0.0118 | 0.8059±0.0130 | 0.8337±0.0125 | 0.7947±0.0108 | 0.8772±0.0070 | 0.8834±0.0078 |