Development of Clinically Informing application based on Recurrent Neural Network (**CIReNN**)

Seng Chan You, MD, [1], Sanghyung Jin[1], Peter R. Rijnbeek, PhD[2], Rae Woong Park, MD, PhD[1,3]

[1]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea;

[2]Erasmus University Medical Center, Rotterdam, The Netherlands;

[2]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea

Background

- Risk prediction for individual patient has been one of the most important themes in clinical research and patient care.
- Traditional approaches usually have used regression-based models, such as the logistic model and the Cox model. These models have been clinically useful and widely accepted because they use a small number of variables which can be easily obtained in clinical practice. However, these models cannot represent more complex relationship among individual predictors and do not model their temporal relationship.
- The Recurrent Neural Network (RNN) model can represent temporal and non-linear relationship among high-dimensional features.

Purpose

- The objective is to build a predictive model based on a recurrent neural network by using temporal features extracted from OMOP-CDM database: **CIReNN** (Clinically Informing application based on Recurrent Neural Network).
- **CIReNN** is expected to facilitate prediction of important clinical events by analyzing flexible and temporal relationships in health care data.

Experiment

- Objective of experiment is to predict 5 year mortality in Korean general elderlies.
- Database
  -National health insurance service (NHIS) sample cohort was converted into OMOP-CDM version 5
  -This database contains consecutive observation for randomly sampled one million general Korean population from 2002 to 2013
- Cohort
  -Target cohort at risk: Subjects who were 65 year or older at index date (2009-01-01)
  -Outcome cohort: Subjects who died between 2009 and 2013 by any cause
  -Among 89,391 target subjects at risk, total of 15,754 (17.6%) developed outcome from 2009 to 2013.
- Train, validation and test set were divided by 0.7, 0.1 and 0.2 ratio, respectively.
- Hyper-parameters of RNN model in the experiment
  -RNN model: single-layer GRU
  -drop-out rate: 0.2
  -activation function: sigmoid
  -optimizer: RMSProp

Model

The whole process of model has 5 steps:

1. Create the risk and outcome cohort by using ATLAS
2. Extract temporal features from the cohort by using temporal_features branch of the feature extraction package
   - The following information is extracted from OMOP-CDM: age, sex, observation, diagnosis history and drug history. For each patient, multi-hot label vectors are generated for representing the patient's medical history as shown in **Fig 1**.

N-dimensional vector

N-dimensional vectors (**sparse array**)

Time span 1 (-1800~1770 day): Fever, Cough[condition], Tylenol[drug] [1,1,0,0,0,…0,0,1,0]

Time span 2 (-1770~1740 day): Pneumonia [condition], Tylenol, Amoxicillin[drug] [0,0,1,0,0,…0,0,1,1]

… …

Time span 50 (-30~ index day): Diabetes mellitus [condition], Insulin[drug] [0,0,0,1,0,…0,0,1,1]

Fig 1. The structure of data containing temporal features, which fed to RNN model

3. Create model settings
   - The model basically use gated recurrent units (GRU), because GRU usually requires less amount of data compared with LSTM.
   - Greedy search algorithm was applied to recommend the best hyper-parameter options by using validation data set.
4. Fitting the model
   - By using Keras with tensorflow backend in R, an RNN model is trained to predict the binary outcome.
5. Evaluate the model

- Acknowledgment : This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea grant number: HI14C3201
- Conflict of interest : none

Experimental result

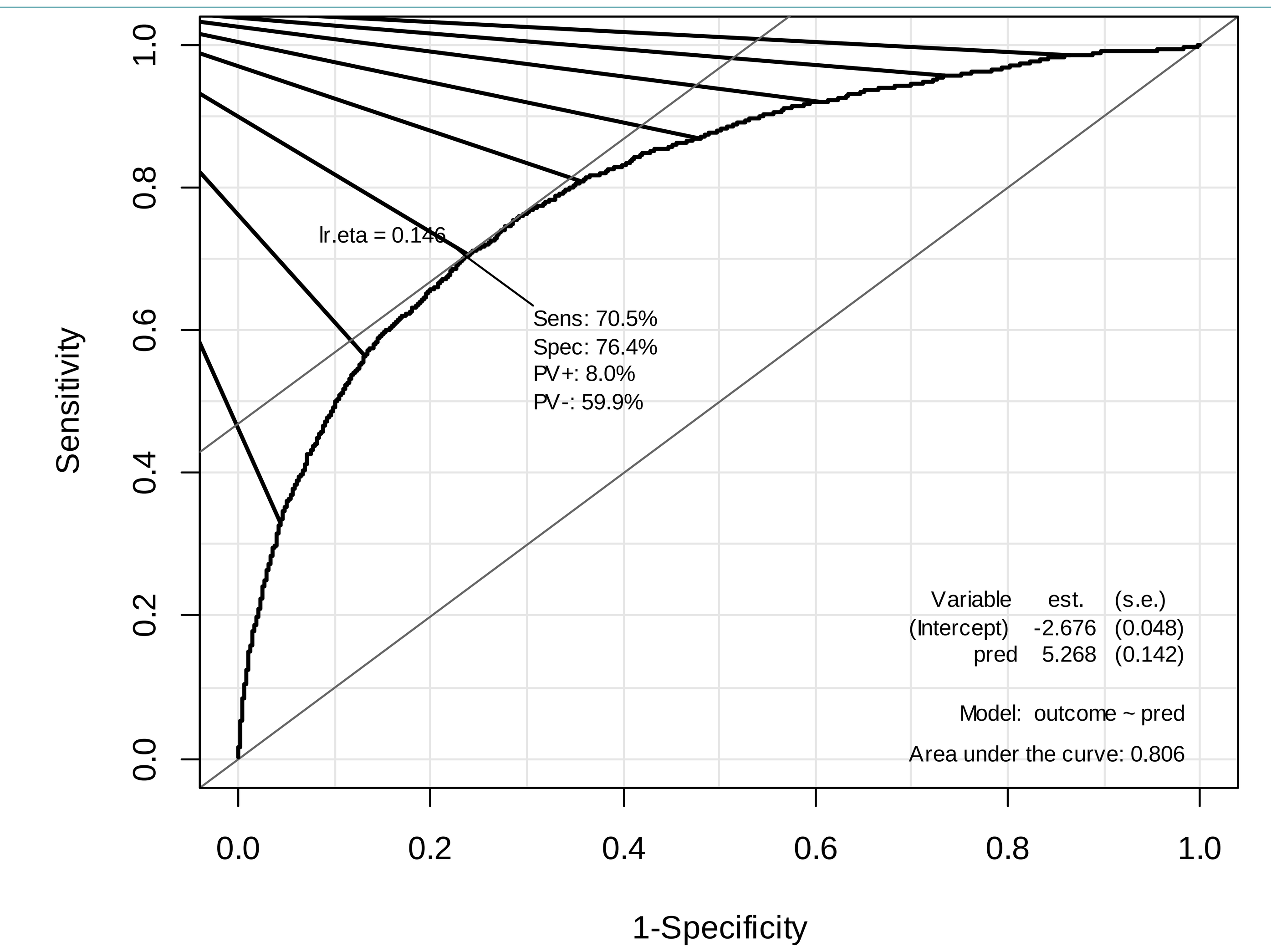

Fig 2. ROC curve for predicting 5-year mortality in elderlies (AUROC = 0.806)

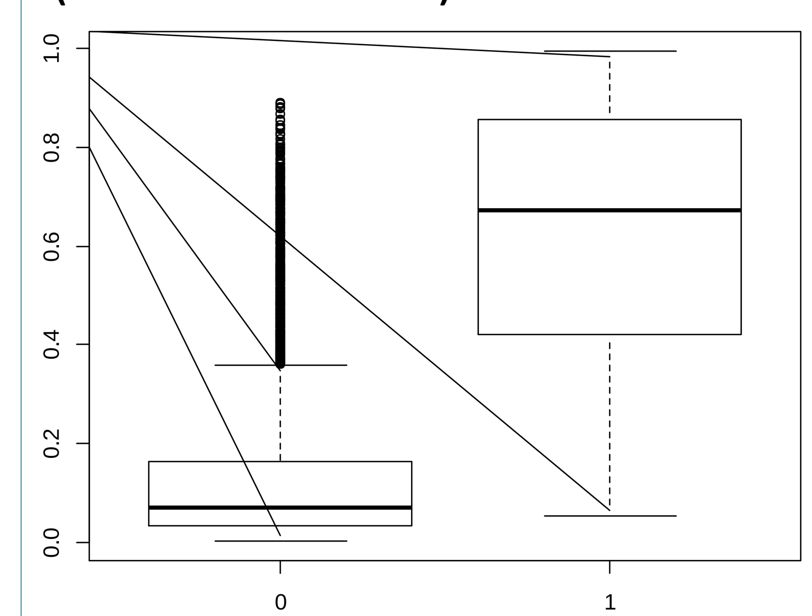

Fig 3. Box plot for prediction value from the model between cohort with and without outcome

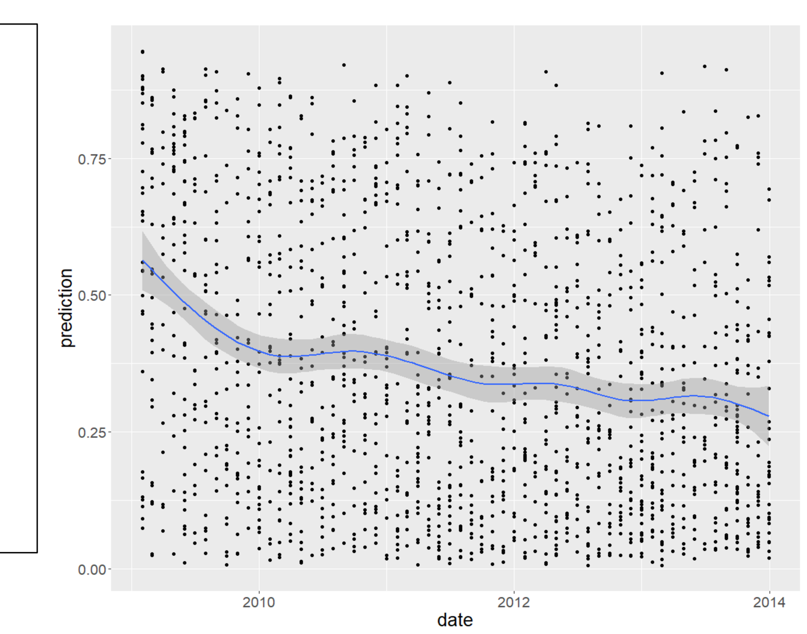

Fig 4. Prediction value for only subjects who died according the death year

- The Area under the ROC curve (AUROC) was 0.8. By taking the optimal cut-off value, positive predictive value and negative predictive value are 0.37 and 0.93, respectively (**Fig 2**).
- The box plot demonstrates that the prediction value from the model has discriminative power to predict 5-year mortality in elderlies (**Fig 3**).
- The predicting power of model decreases as outcome develops later in the target cohort at risk as shown in **Fig 4**.

Conclusion

- We developed a recurrent neural network model, called **CIReNN** for the prediction of future events based on OHDSI platform.
- The feasibility of CIReNN was demonstrated in the experiment, which predicted 5-year mortality in elderlies.
- CIReNN will be integrated into PatientLevelPrediction package in the future