A Machine-Learning Model to Predict Mortality and its Causes Using the National Health Insurance Service National Sample Cohort

Chungsoo Kim, Pharm.D.[1], Seng Chan You, M.D., MS,[2] Rae Woong Park M.D., Ph.D[1,2,3]

[1] Dept. of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, South Korea;
[2] Dept. of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea;
[3] FEEDER-NET(Federated E-Health Big Data for Evidence Renovation Network), Suwon, South Korea

Introduction

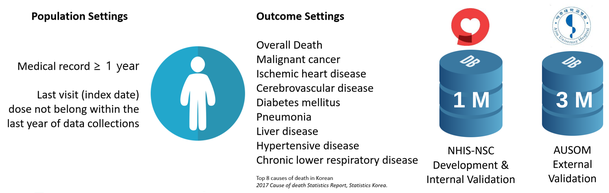
- Mortality data has important value in clinical research but is poorly provided due to privacy concerns and the possibility of social abuse.
- The machine learning can be applied to **multiple classifications**, and a more accurate prediction model can be established through the ensemble technique of multiple algorithms.
- The purpose of this study is **to develop a machine learning model that can predict patient's death and its cause** by using common data model database of National Health Insurance Services' National Sample Cohort (NHIS-NSC) in South Korea.
- All codes are available at https://github.com/ABMI/CauseSpecificMortality

Meanings

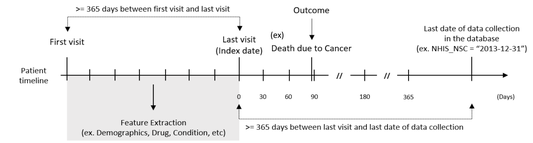
- Using the existing cause of death data, a machine learning model was developed to predict the cause of death.
- We tried to construct a more accurate prediction model through a **stacking ensemble method** of the machine learning algorithm and checked the performance of our models through external validation using Ajou University School Of Medicine (AUSOM) database.
- Since mortality is unbalanced data on outcomes, we tried to characterize the model using indicators such as **AUPRC and F1 score**.
- This study is valuable as a first attempt to develop a cause of death predictive models using claim data linked to the cause of death database. Also in order to predict multiclass outcomes, we proposed a new method in that a stacked model was constructed **using OHDSI's Patient-Level Prediction package**.
- Through our study, we look forward to offering an **alternative to data that lacks the cause of death.**

Methods

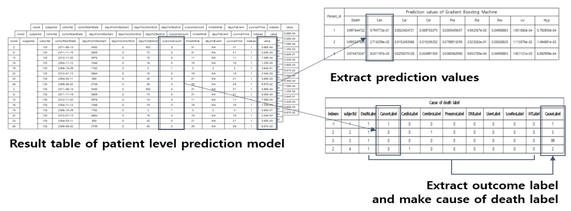
Population Settings and Data sources

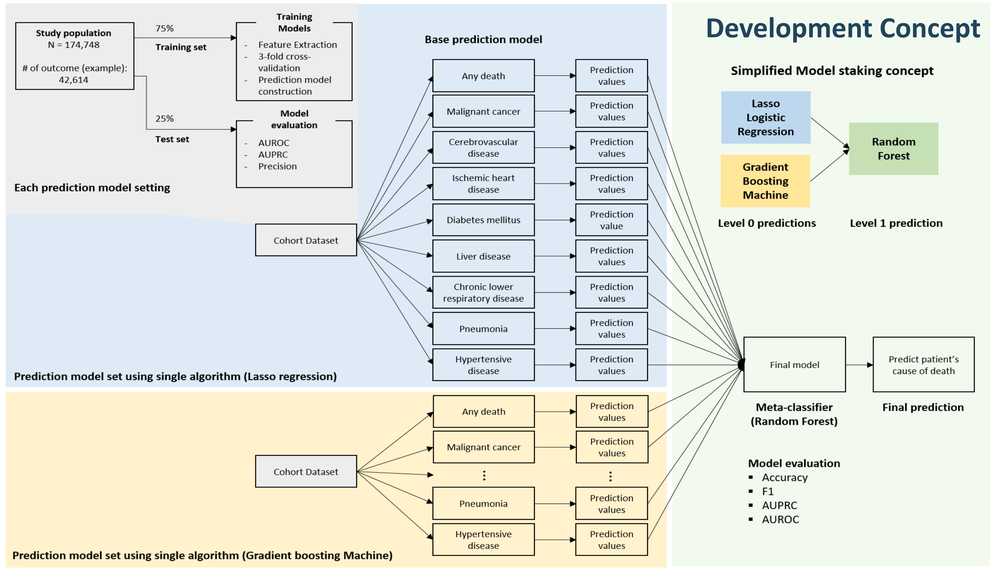


Feature extraction



Applying OHDSI Patient Level Prediction





(Upper Left) Figure 1. Population settings, outcome settings, a schematic view of a patient data extraction.
(Lower Left) Figure 2. Extracting prediction values and outcome labels in patient level prediction package result file.
(Upper Right) Figure 3. Overall prediction model development process.

Results

Result of model development

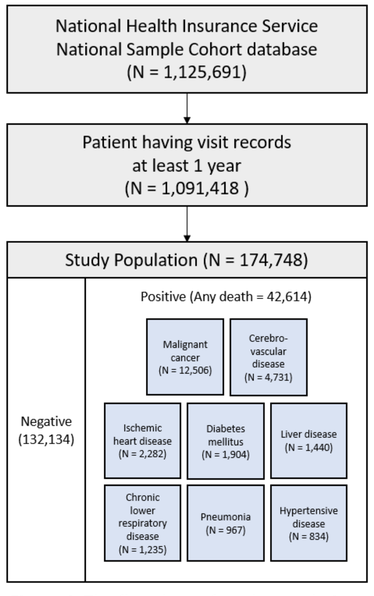


Figure 4. The flowchart of study population.

Table 1. The area under the receiver operating curve in the prediction models

	Lasso logistic regression model					Gradient Boosting Machine model				
Time at risk (days)	30	60	90	180	365	30	60	90	180	365
Causes of death										
Any death	0.9846	0.9862	0.9850	0.9819	0.9810	0.9862	0.9882	0.9867	0.9854	0.9835
Malignant cancer	0.9934	0.9951	0.9951	0.9951	0.9947	0.9941	0.9956	0.9958	0.9947	0.9951
Cerebrovascular disease	0.9804	0.9815	0.9825	0.9805	0.9798	0.9847	0.9838	0.9835	0.9834	0.9795
Ischemic Heart Disease	0.9690	0.9672	0.9655	0.9588	0.9589	0.9710	0.9698	0.9656	0.9636	0.9640
Pneumonia	0.9765	0.9762	0.9666	0.9710	0.9597	0.9718	0.9747	0.9704	0.9721	0.9690
Diabetes Mellitus	0.9822	0.9810	0.9835	0.9817	0.9829	0.9846	0.9855	0.9852	0.9813	0.9822
Liver disease death	0.9919	0.9860	0.9789	0.9784	0.9821	0.9898	0.9861	0.9804	0.9795	0.9792
Chronic lower respiratory disease	0.9895	0.9852	0.9875	0.9868	0.9865	0.9888	0.9868	0.9856	0.9819	0.9852
Hypertensive disease	0.9664	0.9573	0.9590	0.9484	0.9557	0.9546	0.9635	0.9633	0.9597	0.9607

Table 2. The performance of final classifier by time at risk window

TAR (days)	Accuracy	Macro F1	Mean AUPRC	Mean AUROC
30	0.9482	0.6351	0.9725	0.931
60	0.9447	0.6732	0.9937	0.9406
90	0.9277	0.6287	0.9744	0.9234
180	0.9232	0.6115	0.9813	0.9218
365	0.9189	0.6254	0.9838	0.9257

AUPRC : The area under the precision – recall curve, AUROC : The area under the receiver operating characteristics curve

Model Evaluation

- (Table 1) All mortality prediction models showed high **AUROC greater than 0.9**. There was no difference in model performance between the Lasso regression and GBM, between TARs, and between causes of death.
- (Table 2) The performance of the **final classifier** was highest for most indicators when the time at risk window was 60 days.

External validation

- Most death prediction models showed high performance **above AUC 0.9**
- External validations of death from diabetes and hypertensive diseases were not possible due to the lack of the number of patients.
- (Figure 5) The receiver operating characteristics curves of **final classifier** by the time at risk window (30, 60, 90, 180, 365 days).
- (Table 3) The AUROC was highest when the time at risk window was 60 days (**0.8601**), and the AUPRC value was highest when the time at risk window was 365 days (**0.7066**).

Result of External Validation

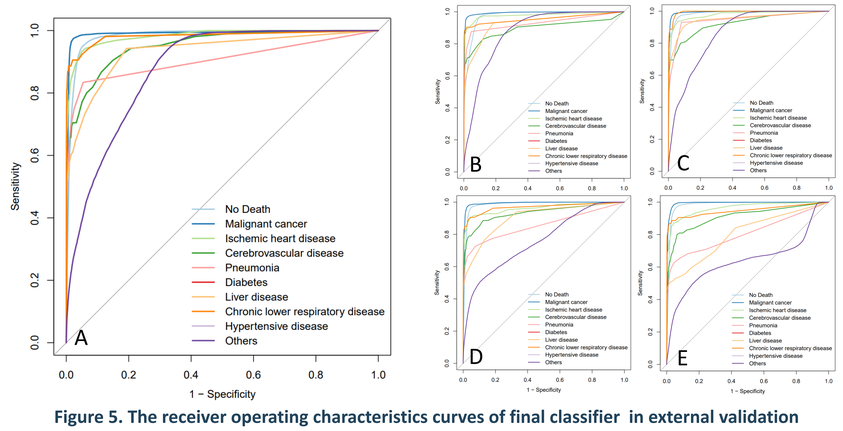


Figure 5. The receiver operating characteristics curves of final classifier in external validation

Table 3. Performance of final classifier by time at risk window in external validation

Graph	TAR (days)	Accuracy	Macro F1	Mean AUPRC	Mean AUROC
A	30	0.9373	0.3380	0.644	0.8409
B	60	0.9235	0.3360	0.6682	0.8601
C	90	0.9237	0.3020	0.6685	0.8520
D	180	0.9177	0.3065	0.6202	0.8409
E	365	0.8299	0.2870	0.7066	0.8299

Conflict of interest : None / **Contact** : Rae Woong Park M.D., PhD., veritas@ajou.ac.kr
Acknowledgements : This work was supported by the Bio Industrial Strategic Technology Development Program (20005021) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea) and a grant from 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: HI16C0992]