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

Chungsoo Kim, PharmD¹, Seng Chan You, MD, MS², Rae Woong Park MD, PhD¹,²,³

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)
Introduction
Death, Cause of death

Death

- Death is clearly of tremendous importance for each individual and also important value in clinical research
- Poorly providing due to privacy concerns and the possibility of social abuse

Cause of death

- All-cause mortality is less sensitive to each disease condition and highly affected by underlying disease
- It can be used for various studies like Global Burden of Disease Study of WHO, Sustainable Development Goals (SDGs)
Attempt to predict death in OHDSI

A study to predict the death by using machine learning

A machine learning model was developed using a Patient-level prediction package provided by OHDSI and external validation was performed

The machine learning model based on OMOP CDM has transferable characteristics that can be easily applied to other institutions.

AUROC : 0.989
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 Sample Cohort in South Korea.
Method
Overall Concept

Patient Level Prediction models

Stacked regularization with PLP outputs

Ensemble model

Final prediction

Predict patient death and its cause
Data Sources

For model develop and internal validation

National Health Insurance Services
National Sample Cohort (NHIS-NSC)
- OMOP CDM
- National Claim database
- No. of patients : 1 millions (Sample)

For External validation

Ajou University School of Medicine (AUSOM)
- OMOP CDM
- Tertiary hospital EMR data
- No. of patients : 3 millions
**Population/outcome settings, Feature extraction**

**Population Settings**
- Medical record ≥ 1 year
- Last visit (index date) dose not belong within the last year of data collections

**Outcome Settings**
- Overall Death
- Malignant cancer
- Ischemic heart disease
- Cerebrovascular disease
- Diabetes mellitus
- Pneumonia
- Liver disease
- Hypertensive disease
- Chronic lower respiratory disease

*Top 8 causes of death in Korea

**Patient Level Prediction**
- Observation Window
- Time-at-risk

**Index date**
- Last visit day

**Time At Risk settings**
- 30, 60, 90, 180, 365 (days)

---

Figure 1. Population settings, outcome settings, a schematic view of a patient data extraction
Applying patient level prediction

Figure 2.
Extracting prediction values and outcome labels in patient level prediction package result file.
Development Concept

Figure 3. Overall prediction model development process
Figure 3. Overall prediction model development process
Development Concept

Figure 3. Overall prediction model development process
Development Concept

Figure 3. Overall prediction model development process
Development Concept

Figure 3. Overall prediction model development process
Figure 3. Overall prediction model development process
Models / covariate settings

- **Model settings**
  - Stacking ensemble model
  - level 0 : Lasso regression
    - Gradient boosting machine
  - level 1 : Random Forest
  - Training : Test = 75 : 25
  - 3-fold cross validation

- **Covariate Settings**
  - 39 covariates
  - Demographics, Condition, Drug, Procedure Observation, Visit Count etc

Figure 4. Simplified model stacking concept
Result

Model development
Flowchart

Database: NHIS-NSC (1M)

Target cohort: 174,748

Outcome cohort (causes of death)
- Any death: 42,614
- Malignant cancer: 12,506
- Cerebrovascular disease: 4,731
- Ischemic heart disease: 2,282
- Diabetes mellitus: 1,904
- Liver disease: 1,440
- Chronic lower respiratory disease: 1,235
- Pneumonia: 967
- Hypertensive disease: 834

Figure 5. The flowchart of study population
PLP results

Time at risk (days) | 30 | 60 | 90 | 180 | 365
---|---|---|---|---|---
Any death
Malignant Cancer
Ischemic heart disease
Cerebrovascular disease
Pneumonia
Diabetes
Liver disease
CLRD
Hypertensive disease

Model 6
Time at risk : 30 days
Outcome : Cancer death
Algorithm : Lasso logistic regression
AUROC : 0.9934

Figure 6. The Receiver Operation Characteristic curves in developed plp models (Lasso logistic regression)
## Table 1. The area under the receiver operating curve in the prediction models (test set)

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Lasso logistic regression model</th>
<th>Gradient Boosting Machine model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>30</td>
<td>60</td>
</tr>
<tr>
<td><strong>Time at risk (days)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Causes of death</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any death</td>
<td>0.9846</td>
<td>0.9862</td>
</tr>
<tr>
<td>Malignant cancer</td>
<td>0.9934</td>
<td>0.9951</td>
</tr>
<tr>
<td>Cerebrovascular disease</td>
<td>0.9804</td>
<td>0.9815</td>
</tr>
<tr>
<td>Ischemic Heart Disease</td>
<td>0.9690</td>
<td>0.9672</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>0.9765</td>
<td>0.9762</td>
</tr>
<tr>
<td>Diabetes Mellitus</td>
<td>0.9822</td>
<td>0.9810</td>
</tr>
<tr>
<td>Liver disease death</td>
<td>0.9919</td>
<td>0.9860</td>
</tr>
<tr>
<td>Chronic lower respiratory disease</td>
<td>0.9895</td>
<td>0.9852</td>
</tr>
<tr>
<td>Hypertensive disease</td>
<td>0.9664</td>
<td>0.9573</td>
</tr>
</tbody>
</table>
Table 2. The performance of final classifier by time at risk window

<table>
<thead>
<tr>
<th>Graph</th>
<th>TAR (days)</th>
<th>Accuracy</th>
<th>Macro F1</th>
<th>Mean AUPRC</th>
<th>Mean AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>30</td>
<td>0.9421</td>
<td>0.6407</td>
<td>0.9736</td>
<td>0.9286</td>
</tr>
<tr>
<td>B</td>
<td>60</td>
<td>0.9389</td>
<td>0.6811</td>
<td>0.9920</td>
<td>0.9347</td>
</tr>
<tr>
<td>C</td>
<td>90</td>
<td>0.9225</td>
<td>0.6394</td>
<td>0.9771</td>
<td>0.9209</td>
</tr>
<tr>
<td>D</td>
<td>180</td>
<td>0.9320</td>
<td>0.6465</td>
<td>0.9810</td>
<td>0.9276</td>
</tr>
<tr>
<td>E</td>
<td>365</td>
<td>0.9265</td>
<td>0.6491</td>
<td>0.9840</td>
<td>0.9294</td>
</tr>
</tbody>
</table>
Final Results – PR curves

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

<table>
<thead>
<tr>
<th>Graph</th>
<th>TAR (days)</th>
<th>Accuracy</th>
<th>Macro F1</th>
<th>Mean AUPRC</th>
<th>Mean AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>30</td>
<td>0.9421</td>
<td>0.6407</td>
<td>0.9736</td>
<td>0.9286</td>
</tr>
<tr>
<td>B</td>
<td>60</td>
<td>0.9389</td>
<td>0.6811</td>
<td>0.9920</td>
<td>0.9347</td>
</tr>
<tr>
<td>C</td>
<td>90</td>
<td>0.9225</td>
<td>0.6394</td>
<td>0.9771</td>
<td>0.9209</td>
</tr>
<tr>
<td>D</td>
<td>180</td>
<td>0.9320</td>
<td>0.6465</td>
<td>0.9810</td>
<td>0.9276</td>
</tr>
<tr>
<td>E</td>
<td>365</td>
<td>0.9265</td>
<td>0.6491</td>
<td>0.9840</td>
<td>0.9294</td>
</tr>
</tbody>
</table>
Result

External validation
Validation Flowchart

Database: AUSOM (3M)

Target cohort: 986,416

Outcome cohort (causes of death)
- Any death: 11,083
- Malignant cancer: 3,064
- Cerebrovascular disease: 110
- Ischemic heart disease: 205
- Liver disease: 485
- Chronic lower respiratory disease: 55
- Pneumonia: 1169
- Diabetes mellitus: 3
- Hypertensive disease: 0

Figure 7. The flowchart of study population in validation dataset
### Validation PLP Results

#### Table 3. The area under the receiver operating curve with external validation set.

<table>
<thead>
<tr>
<th>Causes of death</th>
<th>Lasso logistic regression model</th>
<th>Gradient Boosting Machine model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time at risk (days)</td>
<td>30</td>
</tr>
<tr>
<td>Any death</td>
<td>30</td>
<td>0.9892</td>
</tr>
<tr>
<td>Malignant cancer</td>
<td>60</td>
<td>0.9922</td>
</tr>
<tr>
<td>Cerebrovascular disease</td>
<td>90</td>
<td>0.9189</td>
</tr>
<tr>
<td>Ischemic Heart Disease</td>
<td>180</td>
<td>0.9891</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>365</td>
<td>0.9539</td>
</tr>
<tr>
<td>Chronic lower respiratory disease</td>
<td>30</td>
<td>0.9835</td>
</tr>
<tr>
<td>Liver disease death</td>
<td>60</td>
<td>0.9401</td>
</tr>
<tr>
<td>Diabetes Mellitus</td>
<td>90</td>
<td>-</td>
</tr>
<tr>
<td>Hypertensive disease</td>
<td>180</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>365</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 4. The performance of final classifier by time at risk window

<table>
<thead>
<tr>
<th>Graph</th>
<th>TAR (days)</th>
<th>Accuracy</th>
<th>Macro F1</th>
<th>Mean AUPRC</th>
<th>Mean AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>30</td>
<td>0.9373</td>
<td>0.3380</td>
<td>0.6440</td>
<td>0.8409</td>
</tr>
<tr>
<td>B</td>
<td>60</td>
<td>0.9235</td>
<td>0.3360</td>
<td>0.6682</td>
<td>0.8601</td>
</tr>
<tr>
<td>C</td>
<td>90</td>
<td>0.9237</td>
<td>0.3020</td>
<td>0.6685</td>
<td>0.8520</td>
</tr>
<tr>
<td>D</td>
<td>180</td>
<td>0.9177</td>
<td>0.3065</td>
<td>0.6202</td>
<td>0.8409</td>
</tr>
<tr>
<td>E</td>
<td>365</td>
<td>0.8299</td>
<td>0.2870</td>
<td>0.7066</td>
<td>0.8299</td>
</tr>
</tbody>
</table>
Validation – PR curves

Characteristics

Individual prediction

Stacked prediction

External validation

Table 4. The performance of final classifier by time at risk window

<table>
<thead>
<tr>
<th>Graph</th>
<th>TAR (days)</th>
<th>Accuracy</th>
<th>Macro F1</th>
<th>Mean AUPRC</th>
<th>Mean AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>30</td>
<td>0.9373</td>
<td>0.3380</td>
<td>0.6440</td>
<td>0.8409</td>
</tr>
<tr>
<td>B</td>
<td>60</td>
<td>0.9235</td>
<td>0.3360</td>
<td>0.6682</td>
<td>0.8601</td>
</tr>
<tr>
<td>C</td>
<td>90</td>
<td>0.9237</td>
<td>0.3020</td>
<td>0.6685</td>
<td>0.8520</td>
</tr>
<tr>
<td>D</td>
<td>180</td>
<td>0.9177</td>
<td>0.3065</td>
<td>0.6202</td>
<td>0.8409</td>
</tr>
<tr>
<td>E</td>
<td>365</td>
<td>0.8299</td>
<td>0.2870</td>
<td>0.7066</td>
<td>0.8299</td>
</tr>
</tbody>
</table>
Discussion

• Construction an accurate prediction model through a stacking ensemble method.

• Indicators such as AUPRC and F1 score because mortality is unbalanced data on outcomes.

• First attempt to develop a cause of death predictive models using claim data linked to the cause of death database.

• Proposal for a new method in that a stacked model constructed using OHDSI’s Patient-Level Prediction package.

• We look forward to offering an alternative to data that lacks the cause of death.
Discussion

Developed model evaluation

- 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.

- 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.

- 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).
Conclusions

• Using the existing cause of death data, a machine learning model was developed to predict the cause of death.

• Further study
  - Another external validation (Please Contact)
  - Expand model including other causes of death
  - Model fine tuning
  - Final model selection (other algorithm like GBM, xgboost etc)
  - Death records (Death, Death date, Cause of death) imputation
Thank you

Corresponding Author:
Rae Woong Park, M.D, Ph.D Ajou University School of Medicine
Email: veritas@ajou.ac.kr