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

Chungsoo Kim, PharmD[1], Seng Chan You, MD, MS[2], Rae Woong Park MD, PhD[1,2,3]

[1] Dept of Biomedical Sciences, Ajou Univerisity 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

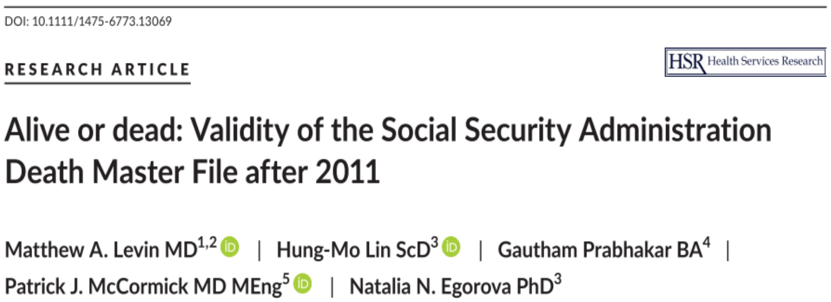
Death, Cause of death

Death

- Death is clearly of tremendous important 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)



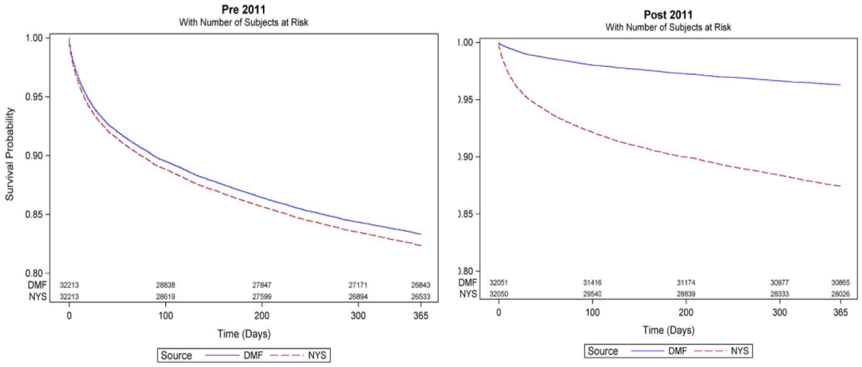
DOI: 10.1111/1475-6773.13069

RESEARCH ARTICLE

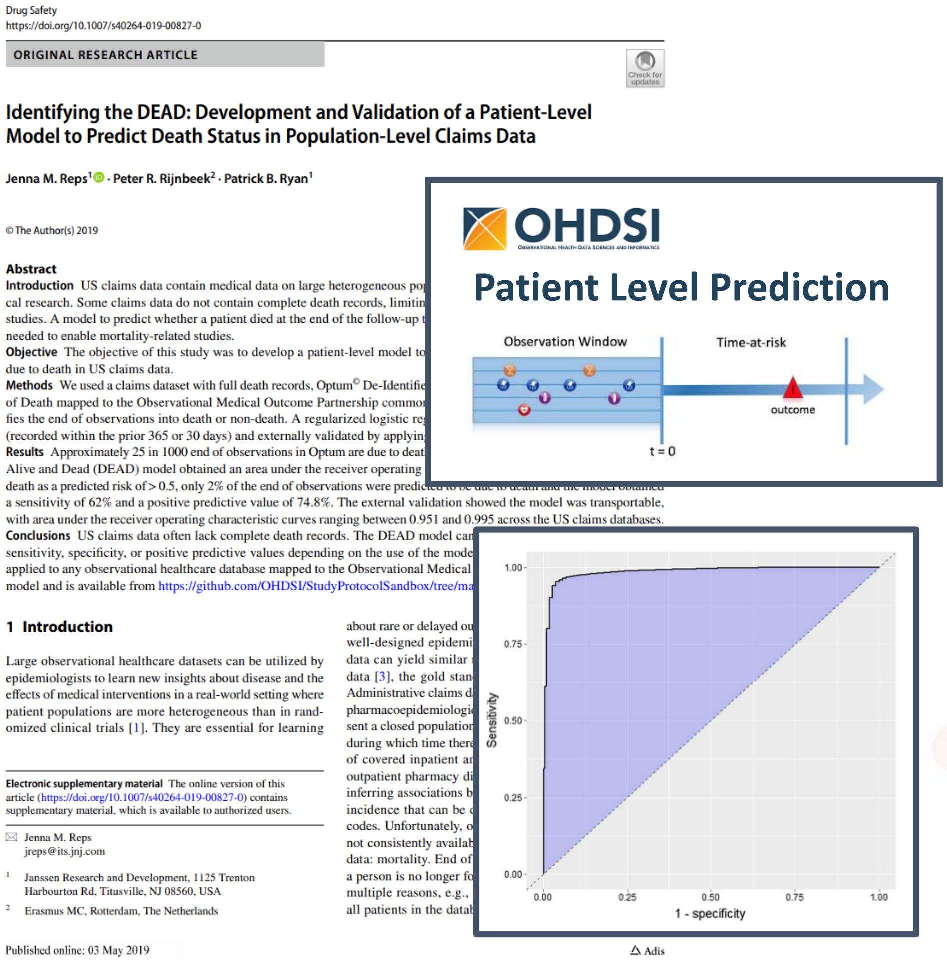
HSR Health Services Research

Alive or dead: Validity of the Social Security Administration Death Master File after 2011

Matthew A. Levin MD[1,2] | Hung-Mo Lin ScD[3] | Gautham Prabhakar BA[4] | Patrick J. McCormick MD MEng[5] | Natalia N. Egorova PhD[3]



Attempt to predict death in OHDSI



Drug Safety
https://doi.org/10.1007/s40264-019-00827-0

ORIGINAL RESEARCH ARTICLE

Check for updates

Identifying the DEAD: Development and Validation of a Patient-Level Model to Predict Death Status in Population-Level Claims Data

Jenna M. Reps[1] · Peter R. Rijnbeek[2] · Patrick B. Ryan[1]

Abstract
Introduction US claims data contain medical data on large heterogeneous po cal research. Some claims data do not contain complete death records, limitin studies. A model to predict whether a patient died at the end of the follow-up t needed to enable mortality-related studies.
Objective The objective of this study was to develop a patient-level model to due to death in US claims data.
Methods We used a claims dataset with full death records, Optum© De-Identifie of Death mapped to the Observational Medical Outcome Partnership commo fies the end of observations into death or non-death. A regularized logistic re (recorded within the prior 365 or 30 days) and externally validated by applyin
Results Approximately 25 in 1000 end of observations in Optum are due to deat Alive and Dead (DEAD) model obtained an area under the receiver operating death as a predicted risk of > 0.5, only 2% of the end of observations were predic a sensitivity of 62% and a positive predictive value of 74.8%. The external validation showed the model was transportable, with area under the receiver operating characteristic curves ranging between 0.951 and 0.995 across the US claims databases.
Conclusions US claims data often lack complete death records. The DEAD model ca sensitivity, specificity, or positive predictive values depending on the use of the mode applied to any observational healthcare database mapped to the Observational Medical model and is available from https://github.com/OHDSI/StudyProtocolSandbox/tree/ma

1 Introduction

Large observational healthcare datasets can be utilized by epidemiologists to learn new insights about disease and the effects of medical interventions in a real-world setting where patient populations are more heterogeneous than in randomized clinical trials [1]. They are essential for learning about rare or delayed ou well-designed epidemi data can yield similar data [3], the gold stan Administrative claims d pharmacoepidemiologi sent a closed population during which time there of covered inpatient ar outpatient pharmacy di inferring associations b incidence that can be c codes. Unfortunately, o not consistently availab data: mortality. End of a person is no longer fo multiple reasons, e.g., all patients in the datab

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40264-019-00827-0) contains supplementary material, which is available to authorized users.

✉ Jenna M. Reps
jreps@its.jnj.com

[1] Janssen Research and Development, 1125 Trenton Harbourton Rd, Titusville, NJ 08560, USA

[2] Erasmus MC, Rotterdam, The Netherlands

Published online: 03 May 2019

- 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.

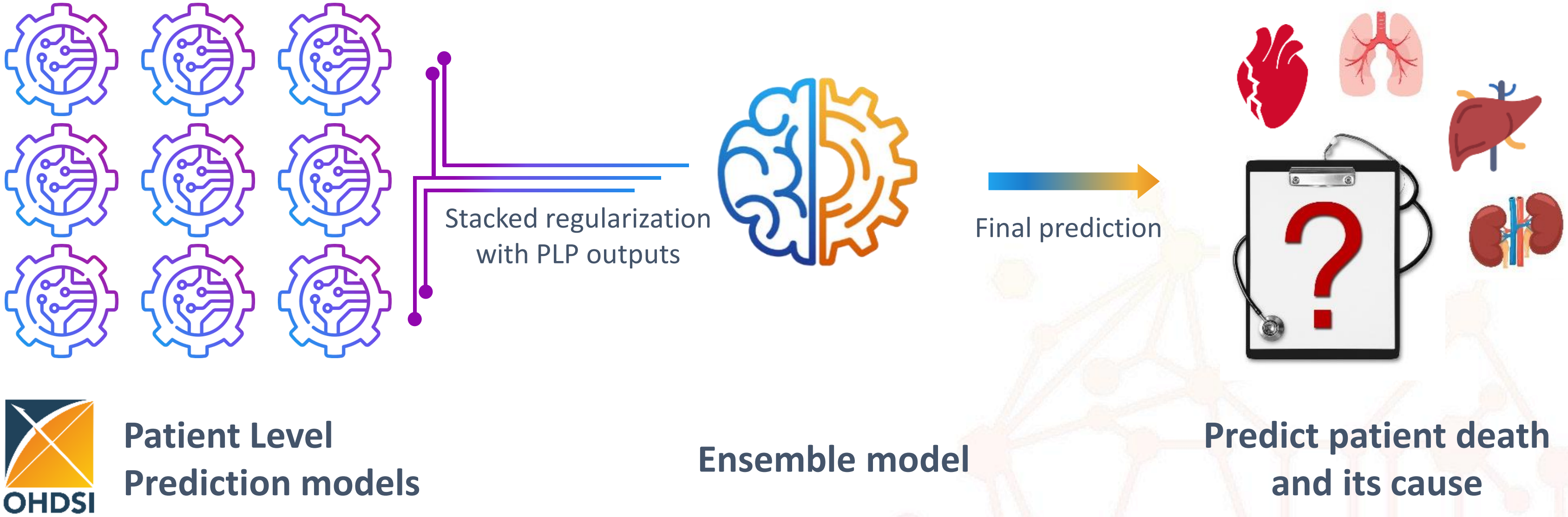
- AUROC : 0.989

Purpose of this study

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



Stacked regularization with PLP outputs

Final prediction

OHDSI

Patient Level Prediction models

Ensemble model

Predict patient death and its cause

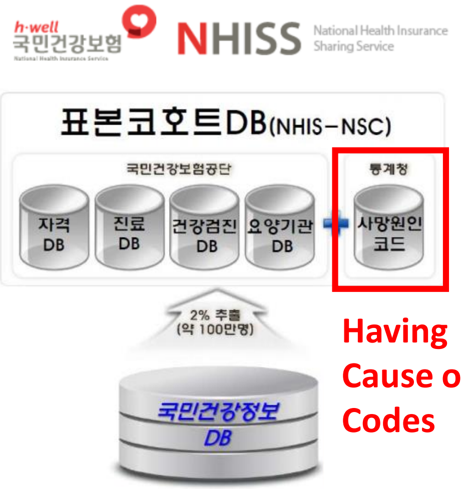
Data Sources

For model develop and internal validation

1 M

National Health Insurance Services National Sample Cohort (NHIS-NSC)

- OMOP CDM
- National Claim database
- No. of patients : 1 millions (Sample)



Having Cause of death Codes

For External validation

3 M

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

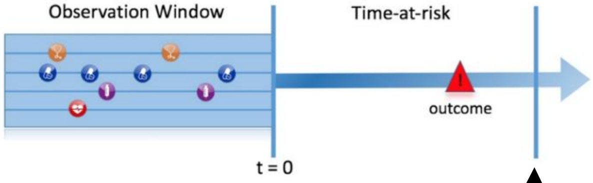
Last visit (index date)
dose not belong within the
last year of data collections

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 Korean
2017 Cause of death Statistics Report, Statistics Korea.

Patient Level Prediction

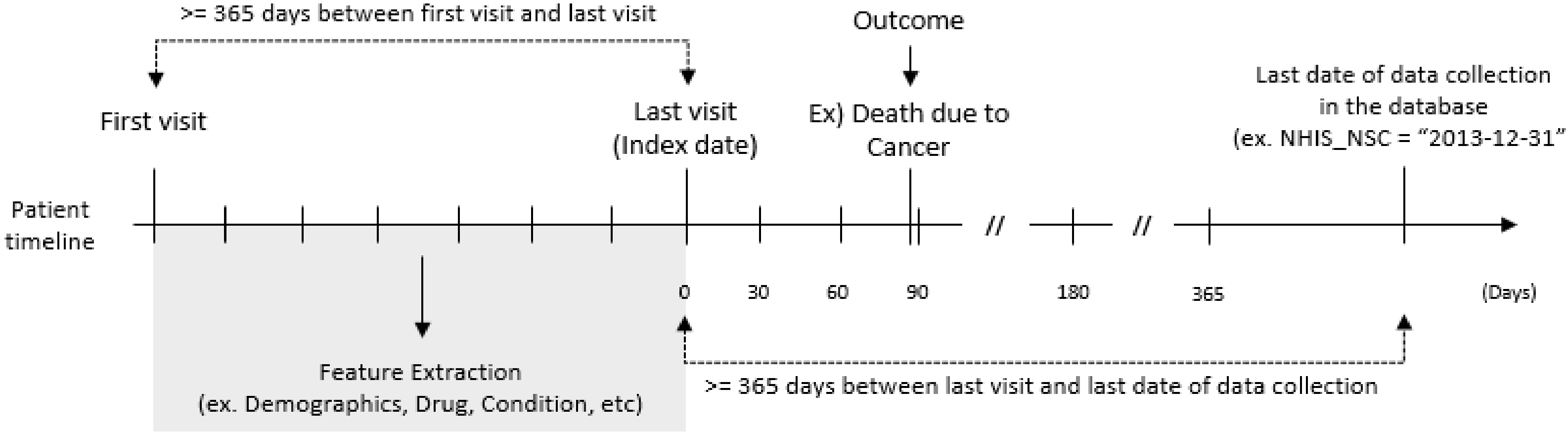


Index date

Last visit day

Time At Risk settings

30, 60, 90, 180, 365 (days)



>= 365 days between first visit and last visit

Outcome

First visit

Last visit
(Index date)

Ex) Death due to
Cancer

Last date of data collection
in the database
(ex. NHIS_NSC = “2013-12-31”)

Patient
timeline

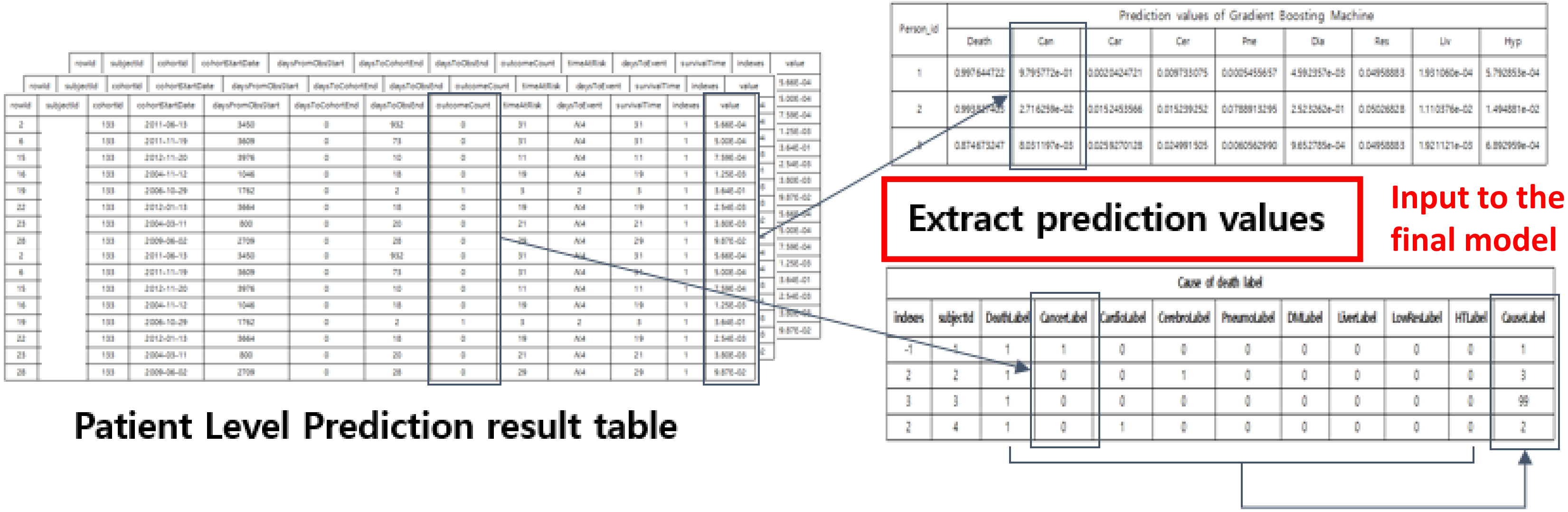
0 30 60 90 180 365 (Days)

Feature Extraction
(ex. Demographics, Drug, Condition, etc)

>= 365 days between last visit and last date of data collection

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

Applying patient level prediction



rowId	subjectId	cohortId	cohortStartDate	daysFromObsStart	daysToCohortEnd	daysToObsEnd	outcomeCount	timeAtRisk	daysToEvent	survivalTime	indexes	value
2		133	2011-06-13	3450	0	932	0	31	NA	31	1	5.66E-04
6		133	2011-11-19	3609	0	73	0	31	NA	31	1	5.00E-04
15		133	2012-11-20	3976	0	10	0	11	NA	11	1	7.59E-04
16		133	2004-11-12	1046	0	18	0	19	NA	19	1	1.25E-03
19		133	2006-10-29	1762	0	2	1	3	2	3	1	3.64E-01
22		133	2012-01-13	3664	0	18	0	19	NA	19	1	2.54E-03
23		133	2004-03-11	800	0	20	0	21	NA	21	1	3.80E-03
28		133	2009-06-02	2709	0	28	0	29	NA	29	1	9.87E-02
2		133	2011-06-13	3450	0	932	0	31	NA	31	1	5.66E-04
6		133	2011-11-19	3609	0	73	0	31	NA	31	1	5.00E-04
15		133	2012-11-20	3976	0	10	0	11	NA	11	1	7.59E-04
16		133	2004-11-12	1046	0	18	0	19	NA	19	1	1.25E-03
19		133	2006-10-29	1762	0	2	1	3	2	3	1	3.64E-01
22		133	2012-01-13	3664	0	18	0	19	NA	19	1	2.54E-03
23		133	2004-03-11	800	0	20	0	21	NA	21	1	3.80E-03
28		133	2009-06-02	2709	0	28	0	29	NA	29	1	9.87E-02

Patient Level Prediction result table

Person_id	Prediction values of Gradient Boosting Machine								
	Death	Can	Car	Cer	Pne	Dia	Res	Liv	Hyp
1	0.997644722	9.795772e-01	0.0020424721	0.009733075	0.0005455657	4.592357e-03	0.04958883	1.951060e-04	5.792853e-04
2	0.993837485	2.716259e-02	0.0152453566	0.015239252	0.0788913295	2.525262e-01	0.05026828	1.110376e-02	1.494881e-02
3	0.874673247	8.051197e-03	0.0258270128	0.024991505	0.0080562990	9.652785e-04	0.04958883	1.921121e-03	6.892959e-04

Extract prediction values

Input to the final model

Cause of death label											
indexes	subjectId	DeathLabel	CancerLabel	CardioLabel	CerebroLabel	PneumoLabel	DMLabel	LiverLabel	LowResLabel	HTLabel	CauseLabel
-1	1	1	1	0	0	0	0	0	0	0	1
2	2	1	0	0	1	0	0	0	0	0	3
3	3	1	0	0	0	0	0	0	0	0	99
2	4	1	0	1	0	0	0	0	0	0	2

Extract outcome label
and make cause of death label

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

Development Concept

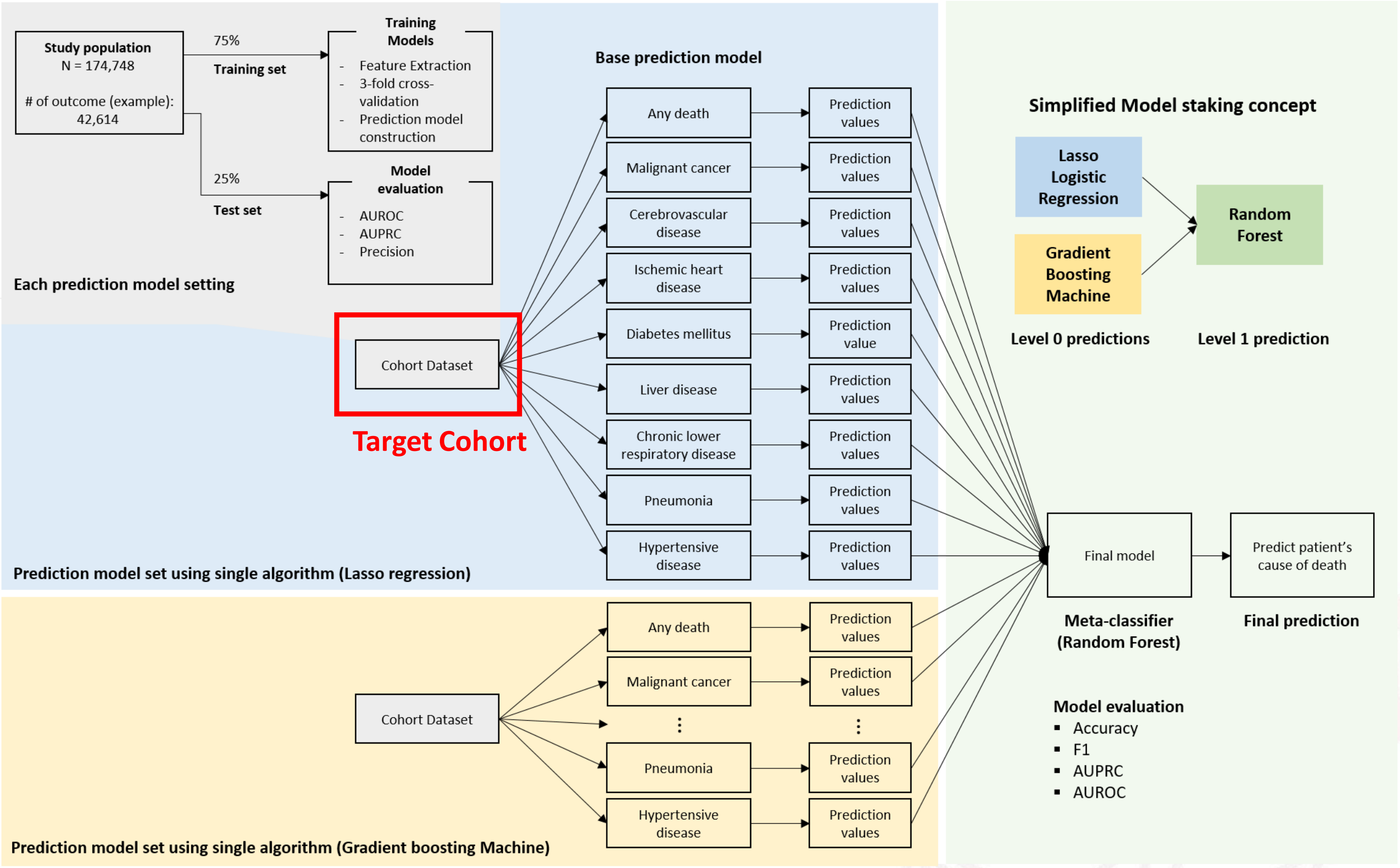


Figure 3. Overall prediction model development process

Development Concept

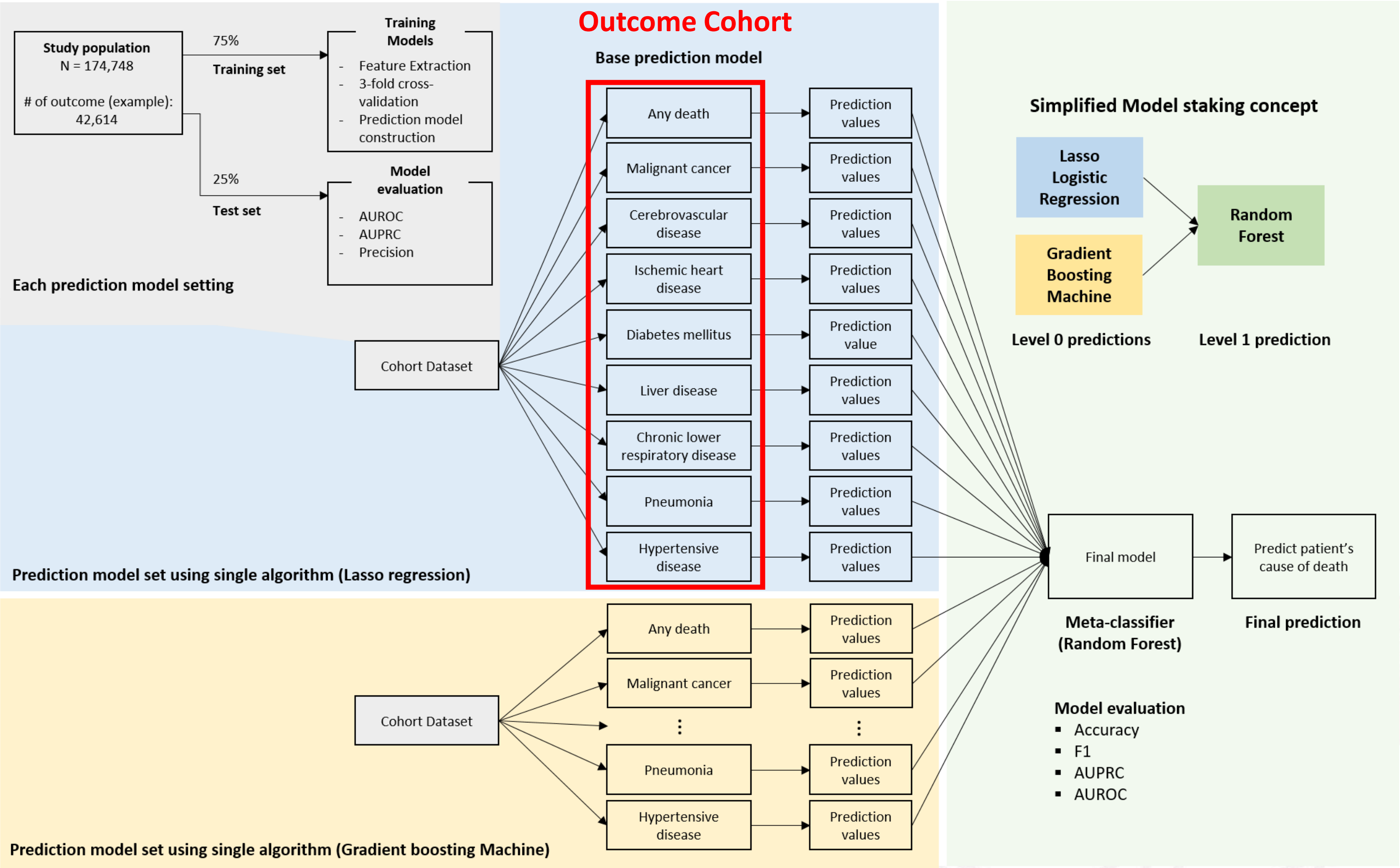


Figure 3. Overall prediction model development process

Development Concept

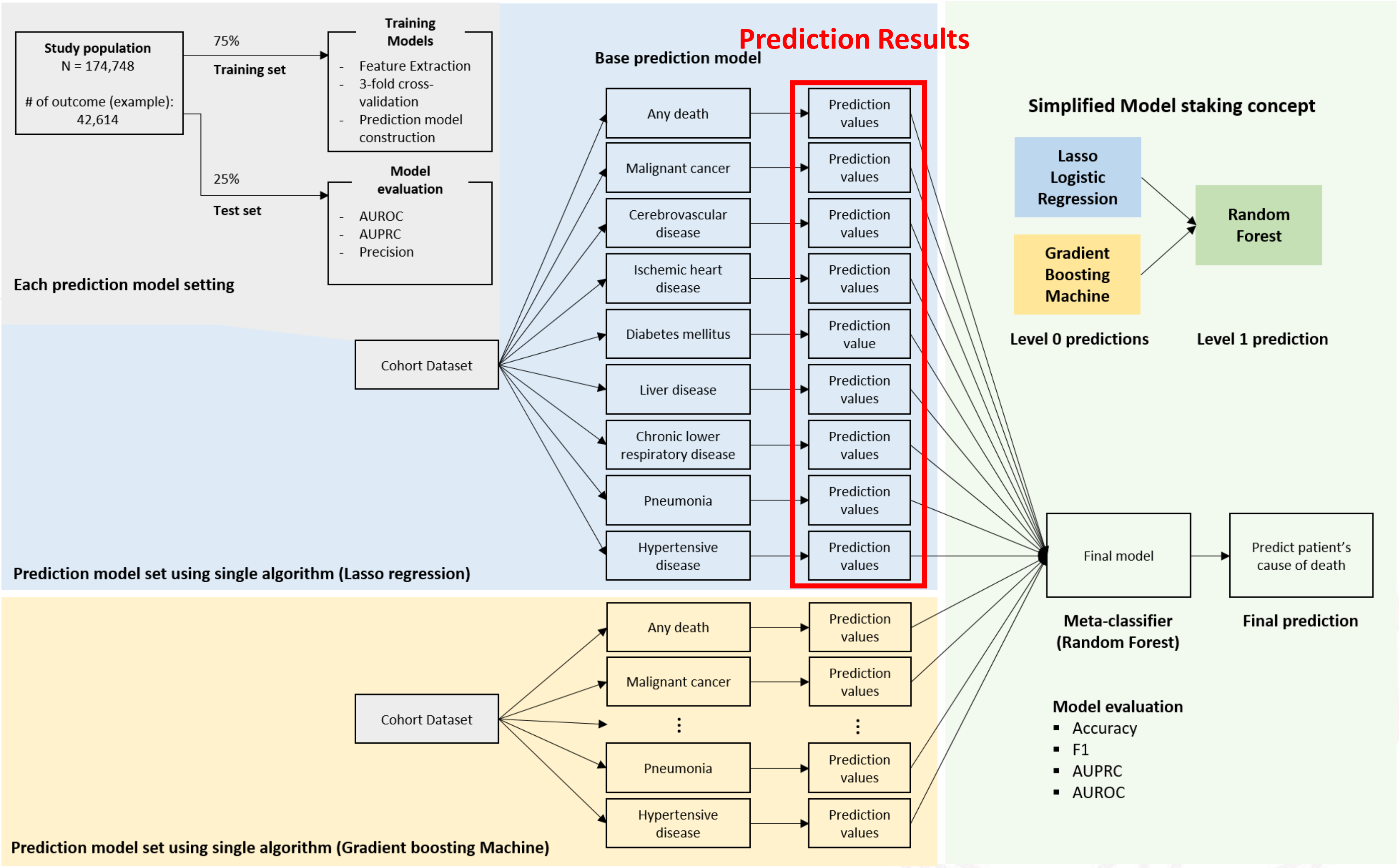


Figure 3. Overall prediction model development process

Development Concept

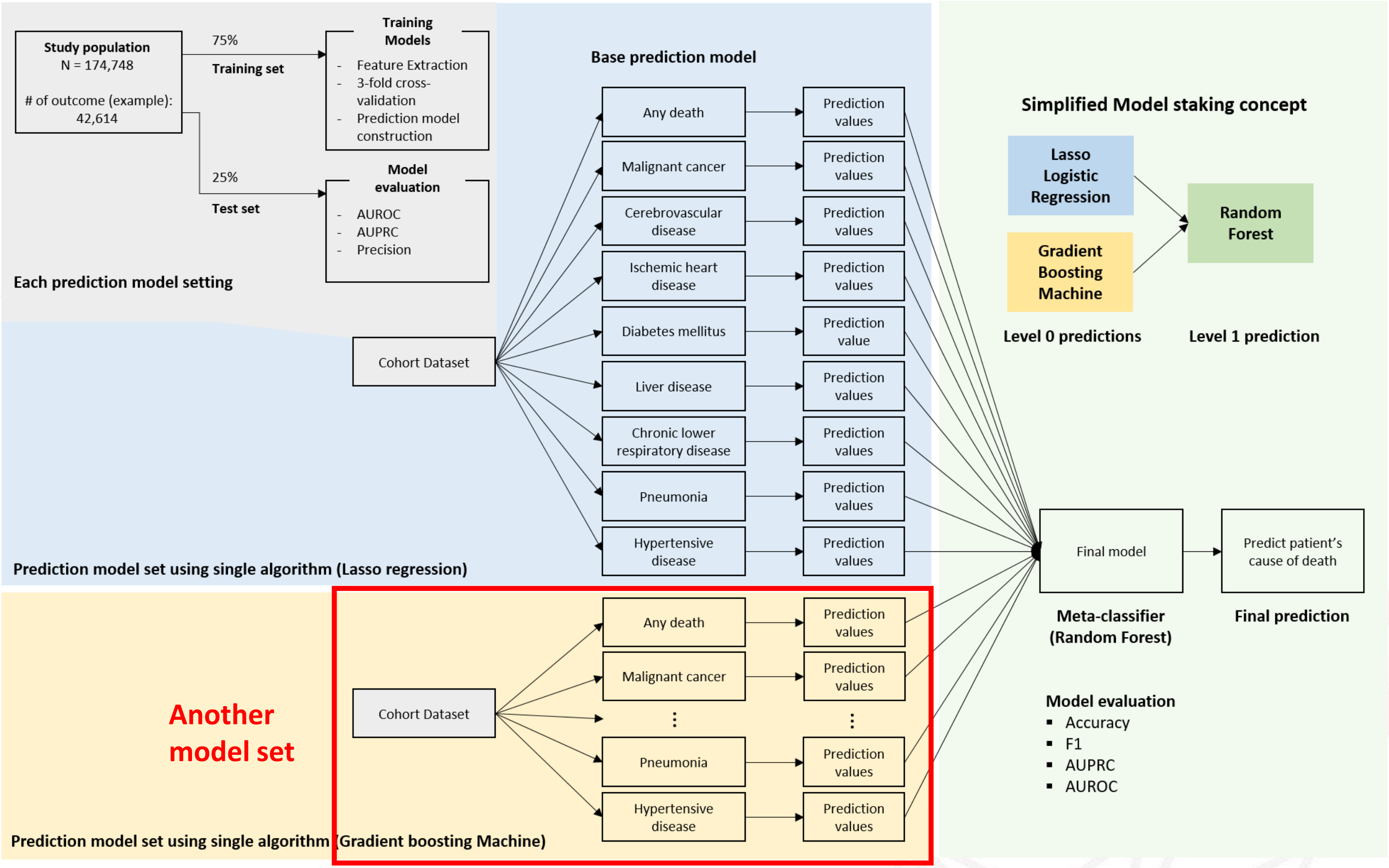
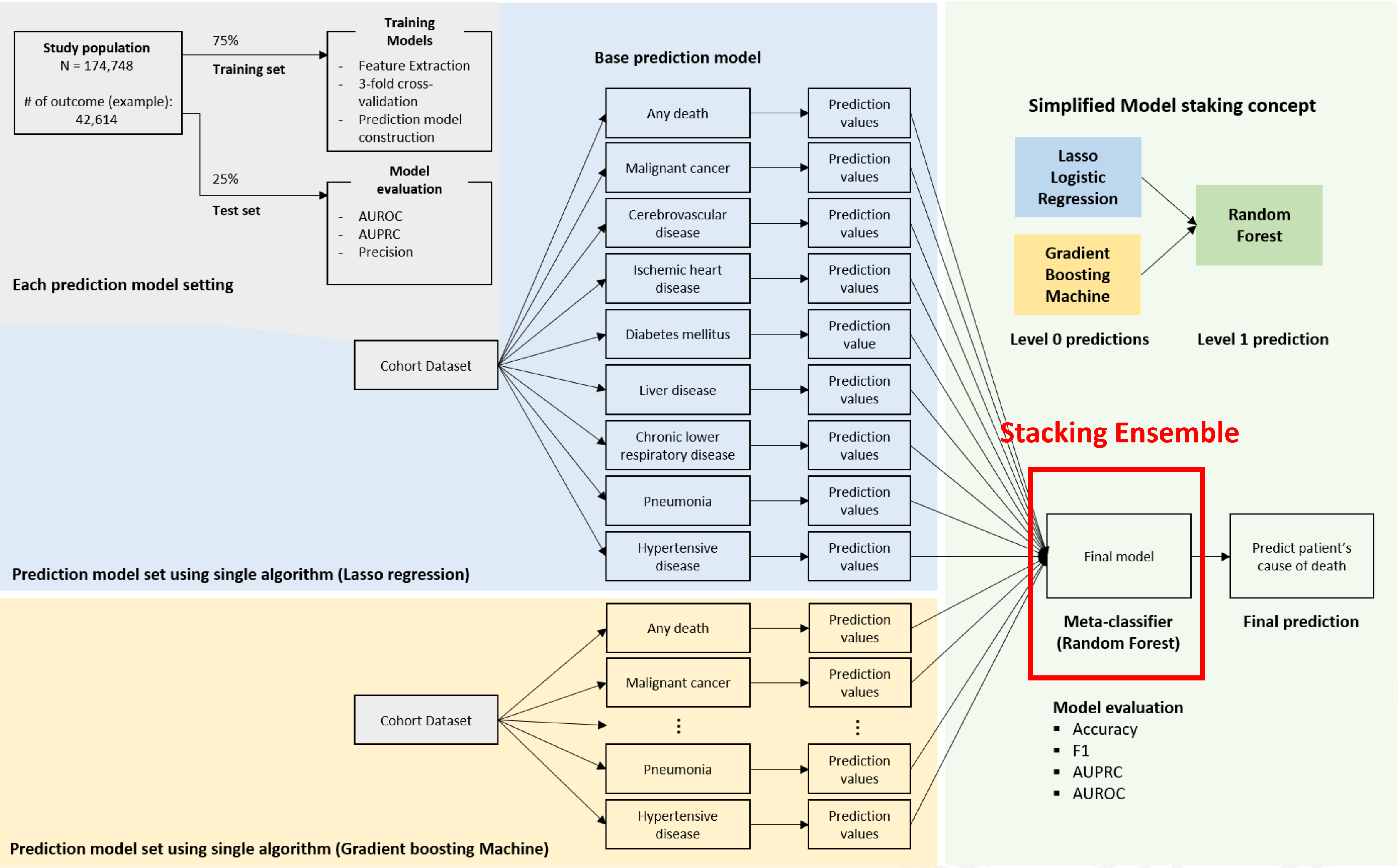


Figure 3. Overall prediction model development process

Development Concept



Study population
N = 174,748

# of outcome (example):
42,614

75%

Training set

25%

Test set

Training Models
- Feature Extraction
- 3-fold cross-validation
- Prediction model construction

Model evaluation
- AUROC
- AUPRC
- Precision

Each prediction model setting

Base prediction model

Cohort Dataset

Any death → Prediction values

Malignant cancer → Prediction values

Cerebrovascular disease → Prediction values

Ischemic heart disease → Prediction values

Diabetes mellitus → Prediction value

Liver disease → Prediction values

Chronic lower respiratory disease → Prediction values

Pneumonia → Prediction values

Hypertensive disease → Prediction values

Prediction model set using single algorithm (Lasso regression)

Cohort Dataset

Any death → Prediction values

Malignant cancer → Prediction values

⋮

Pneumonia → Prediction values

Hypertensive disease → Prediction values

Prediction model set using single algorithm (Gradient boosting Machine)

Simplified Model staking concept

Lasso Logistic Regression

Gradient Boosting Machine

Random Forest

Level 0 predictions

Level 1 prediction

Stacking Ensemble

Final model

Meta-classifier (Random Forest)

Predict patient's cause of death

Final prediction

Model evaluation
- Accuracy
- F1
- AUPRC
- AUROC

Figure 3. Overall prediction model development process

Development Concept

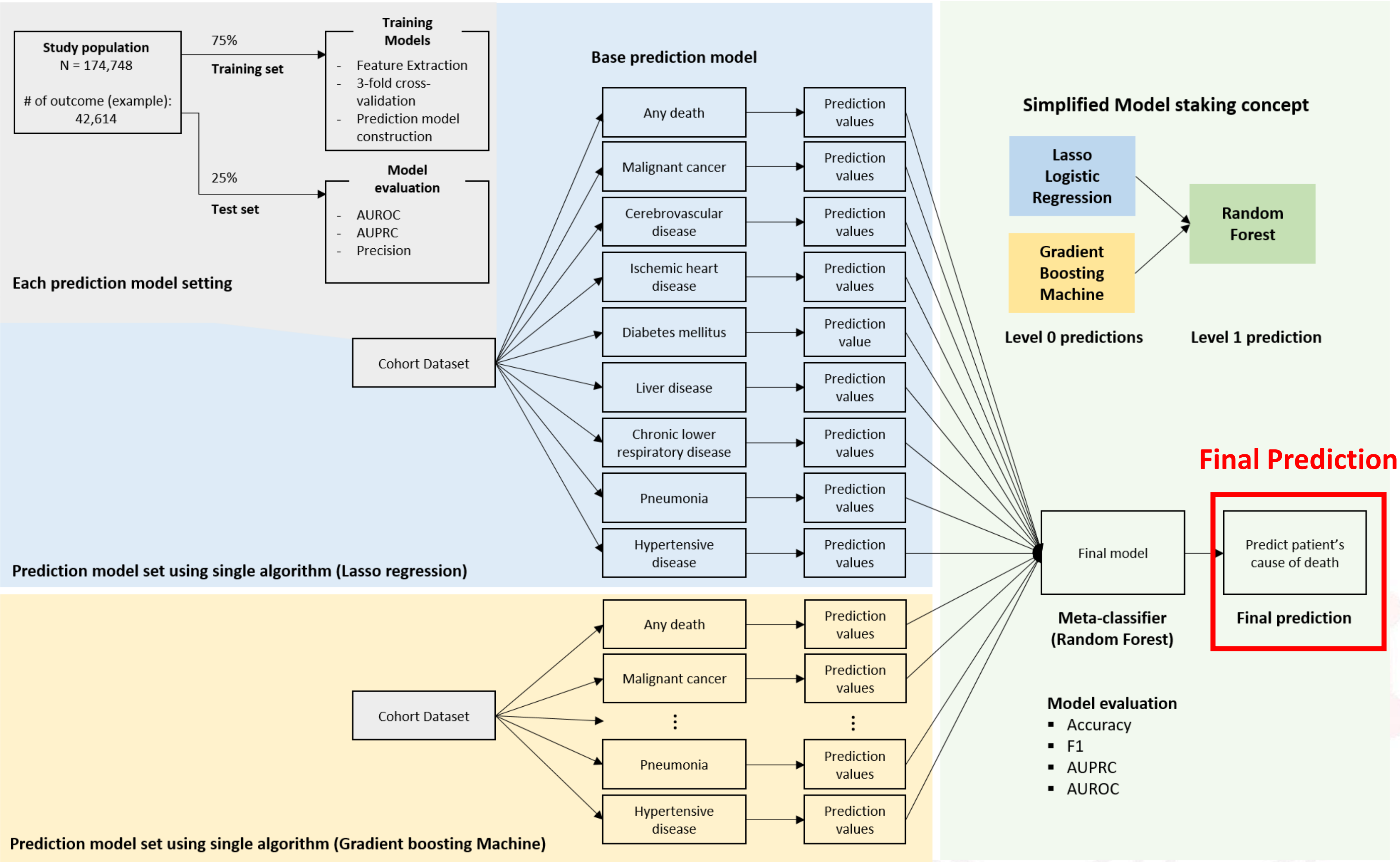


Figure 3. Overall prediction model development process

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

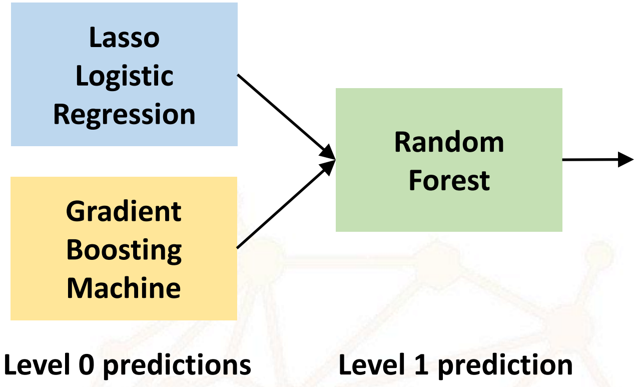


Figure 4. Simplified model stacking concept

- **Covariate Settings**

  - 39 covariates

  - Demographics, Condition, Drug, Procedure Observation, Visit Count etc

Result

Model development

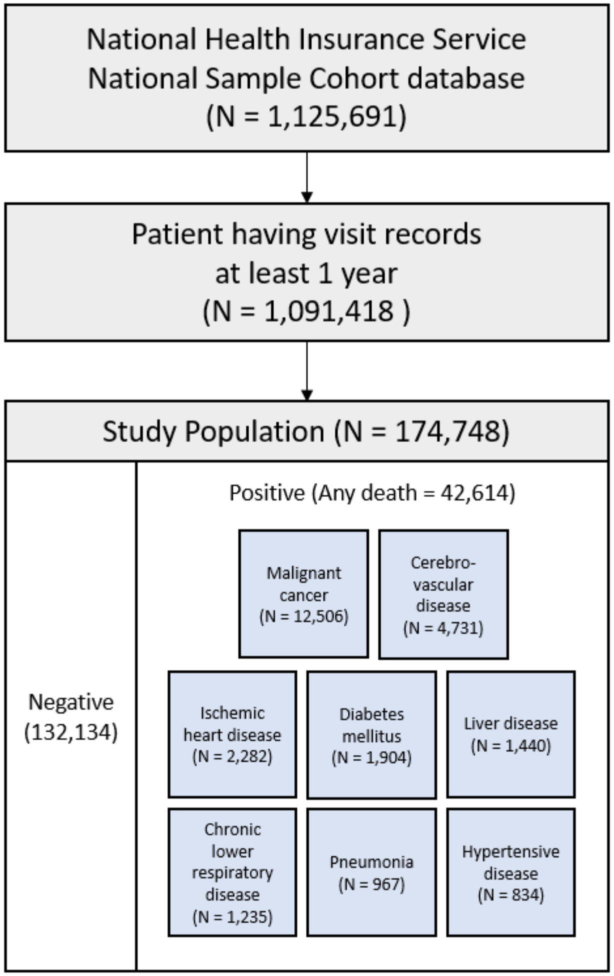
Flowchart

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



National Health Insurance Service National Sample Cohort database (N = 1,125,691)

Patient having visit records at least 1 year (N = 1,091,418)

Study Population (N = 174,748)

Negative (132,134)

Positive (Any death = 42,614)

Malignant cancer (N = 12,506)

Cerebro-vascular disease (N = 4,731)

Ischemic heart disease (N = 2,282)

Diabetes mellitus (N = 1,904)

Liver disease (N = 1,440)

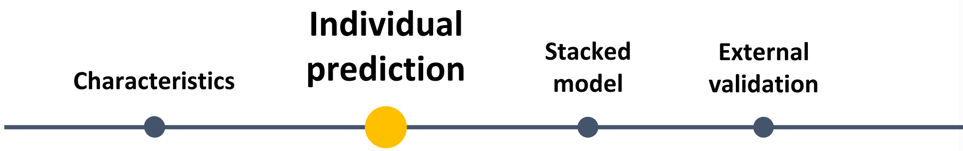
Chronic lower respiratory disease (N = 1,235)

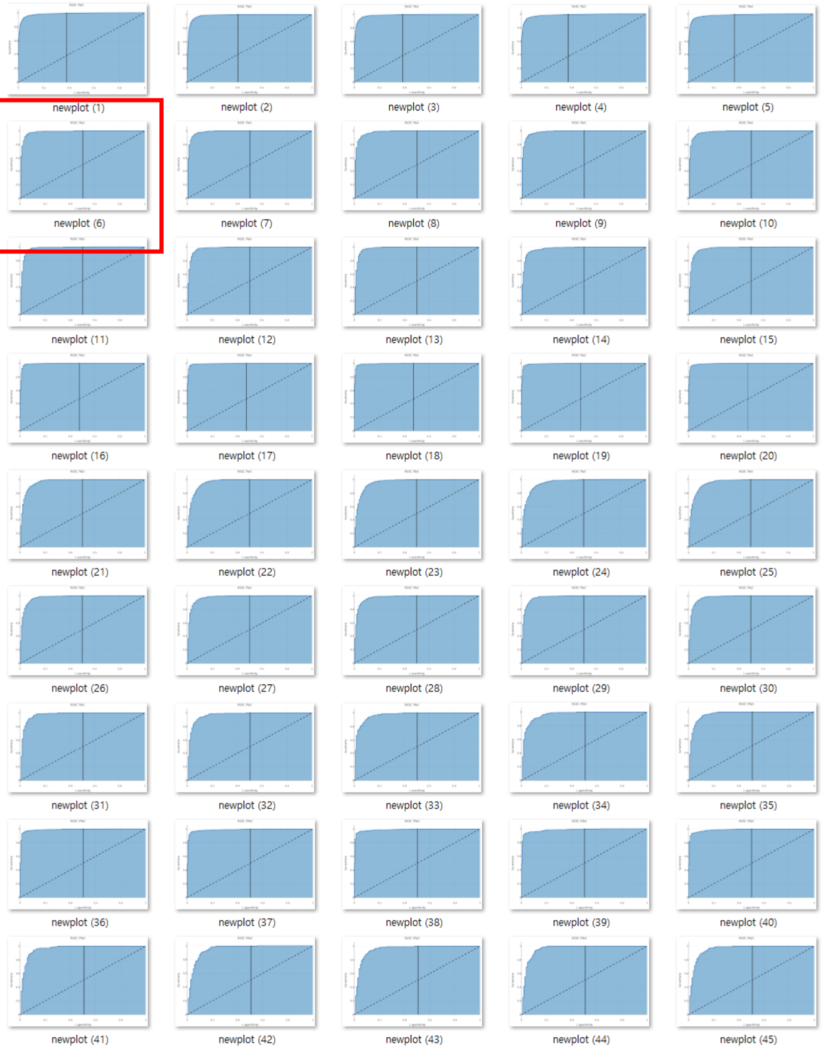
Pneumonia (N = 967)

Hypertensive disease (N = 834)

Figure 5. The flowchart of study population

PLP results





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)

PLP Results

Table 1. The area under the receiver operating curve in the prediction models (test set)

	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

Final Results – ROC curves

Characteristics · Individual prediction · **Stacked prediction** · External validation

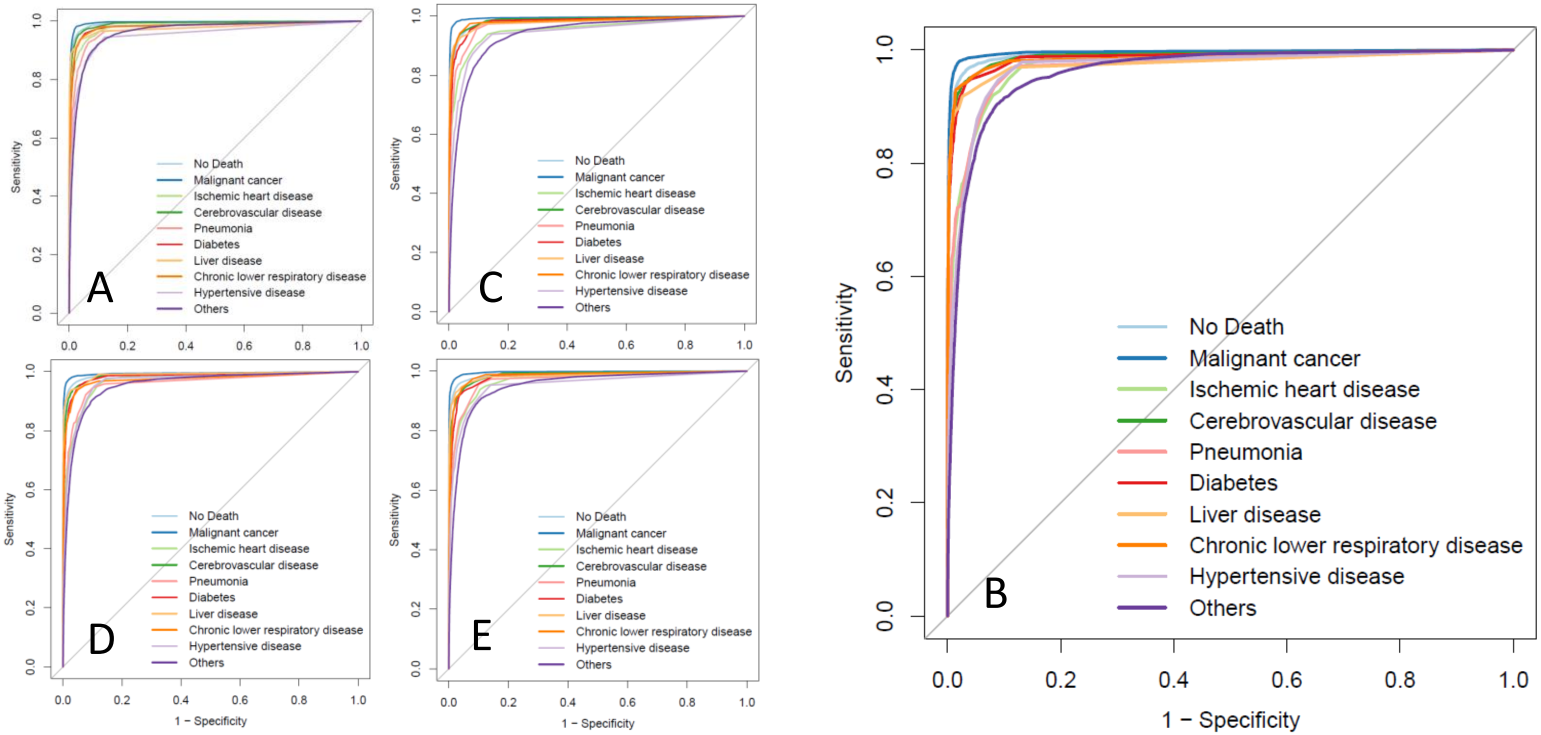
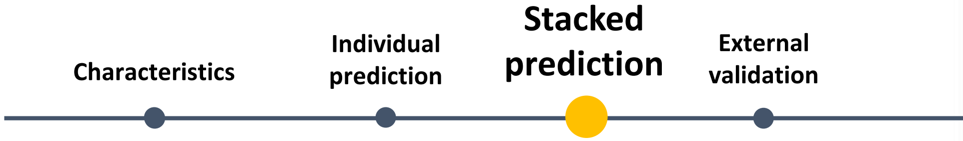


Table2. The performance of final classifier by time at risk window 190911. revised

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

Final Results – PR curves



Characteristics

Individual prediction

Stacked prediction

External validation

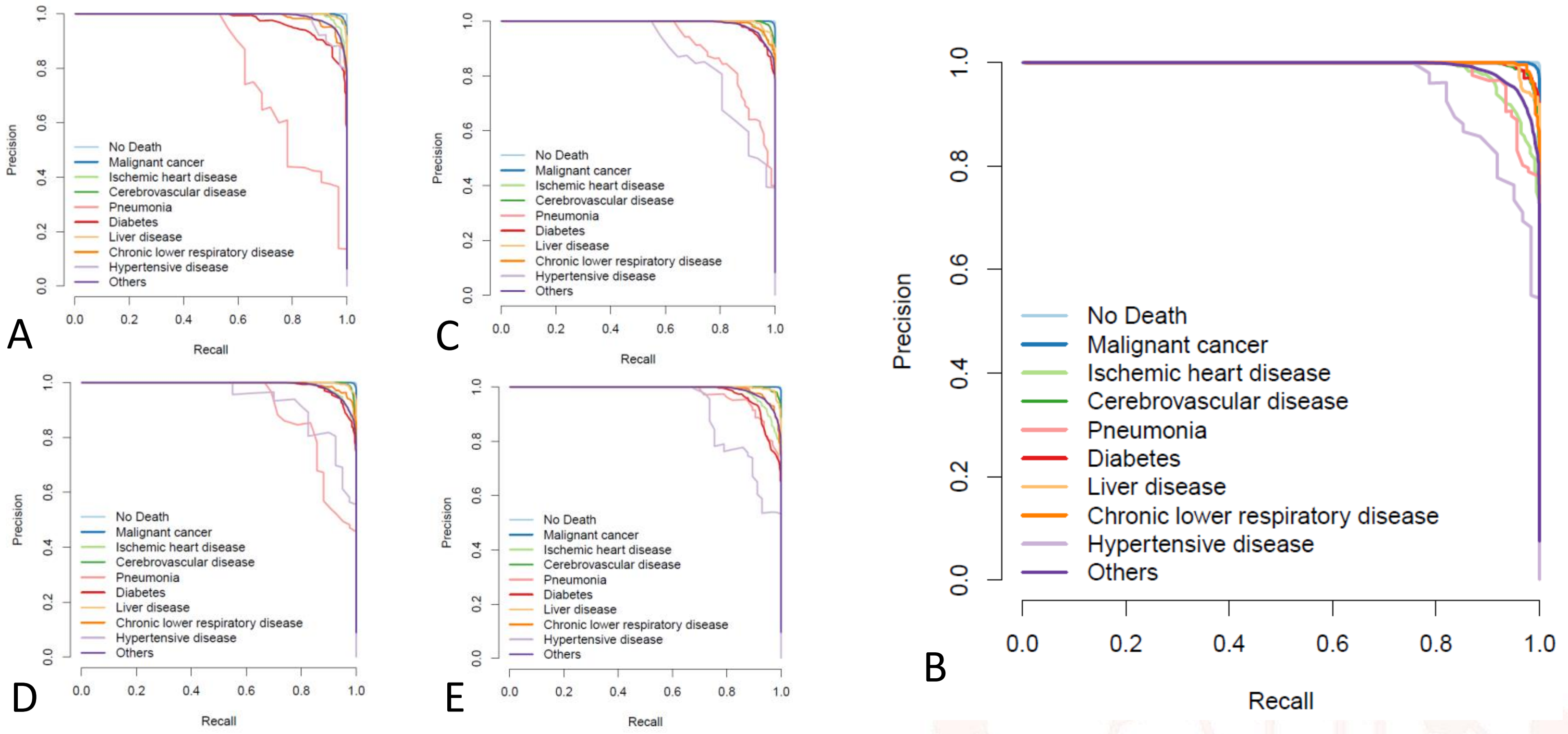


Table2. The performance of final classifier by time at risk window

190911. revised

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

Result

External validation

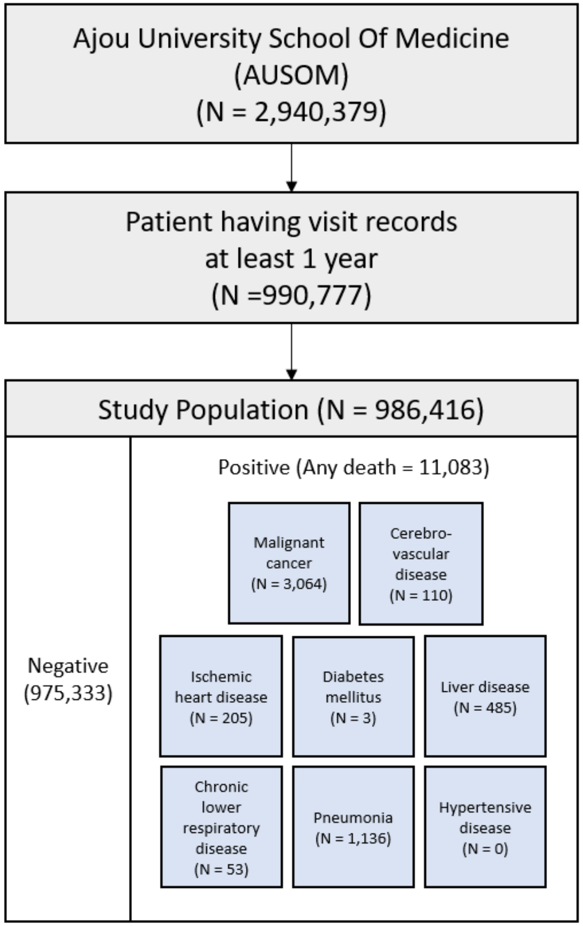
Validation Flowchart

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



Ajou University School Of Medicine (AUSOM) (N = 2,940,379)

Patient having visit records at least 1 year (N =990,777)

Study Population (N = 986,416)

Negative (975,333)

Positive (Any death = 11,083)

Malignant cancer (N = 3,064)

Cerebro-vascular disease (N = 110)

Ischemic heart disease (N = 205)

Diabetes mellitus (N = 3)

Liver disease (N = 485)

Chronic lower respiratory disease (N = 53)

Pneumonia (N = 1,136)

Hypertensive disease (N = 0)

Figure 7. The flowchart of study population in validation dataset

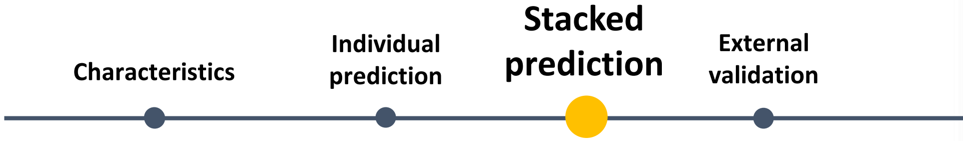
Validation PLP Results

Characteristics | Individual predictions | Stacked model | External validation

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

	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	0.9892	0.9896	0.9891	0.9899	0.9889	0.9875	0.9889	0.9895	0.9897	0.9907
Malignant cancer	0.9922	0.9934	0.9894	0.9924	0.9933	0.9943	0.9929	0.9931	0.9925	0.9913
Cerebrovascular disease	0.9189	0.8934	0.9225	0.8999	0.8669	0.9734	0.9739	0.9665	0.9582	0.9361
Ischemic Heart Disease	0.9891	0.9795	0.9854	0.9607	0.9852	0.9897	0.9864	0.9827	0.9802	0.9757
Pneumonia	0.9539	0.9350	0.9241	0.9345	0.9349	0.9833	0.9777	0.9667	0.9605	0.9409
Chronic lower respiratory disease	0.9835	0.9849	0.9872	0.9821	0.9828	0.9974	0.9976	0.9977	0.9979	0.9944
Liver disease death	0.9401	0.8819	0.8895	0.8974	0.8805	0.9937	0.9950	0.9914	0.9877	0.9654
Diabetes Mellitus	-	-	-	-	-	-	-	-	-	-
Hypertensive disease	-	-	-	-	-	-	-	-	-	-

Validation – ROC curves



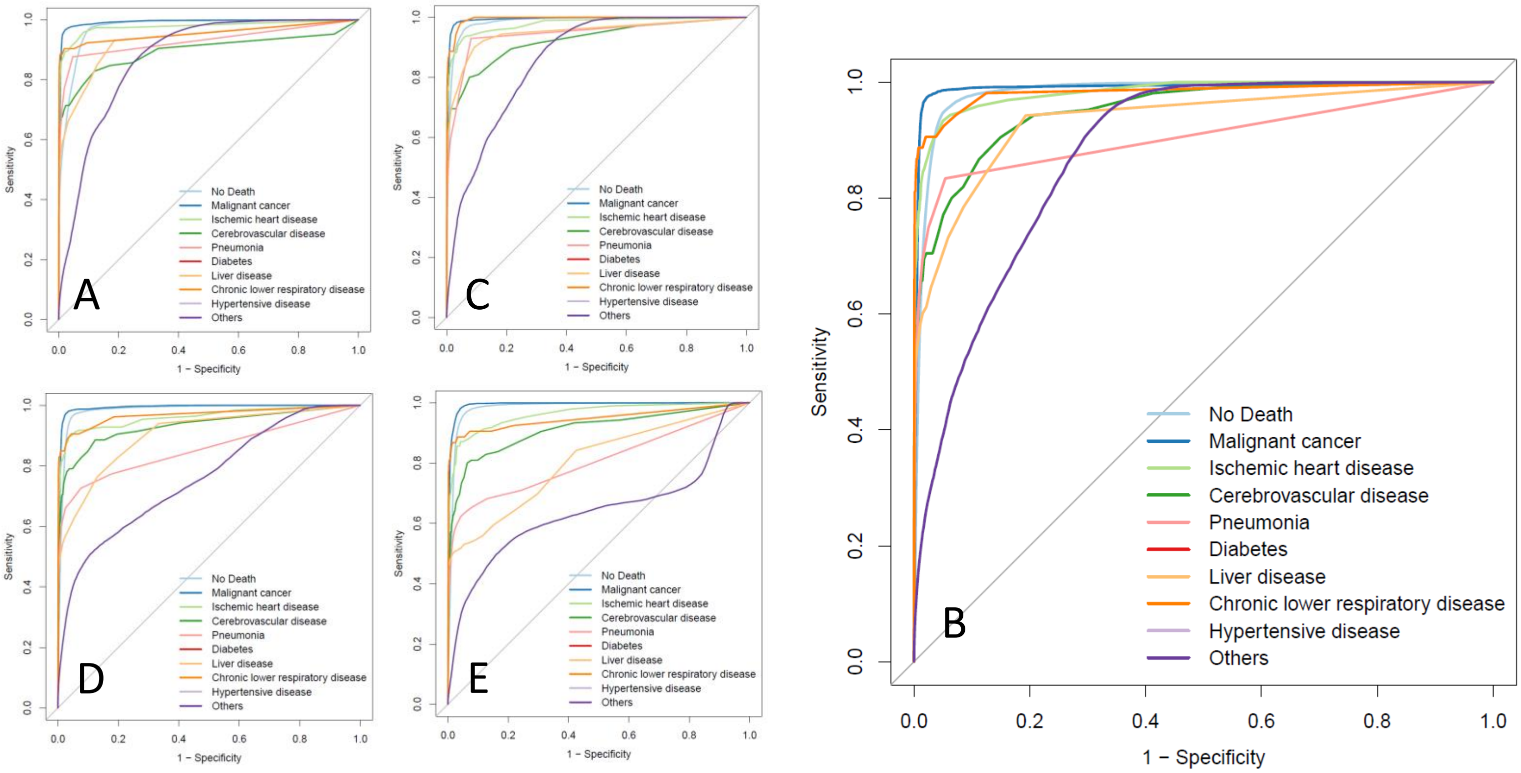
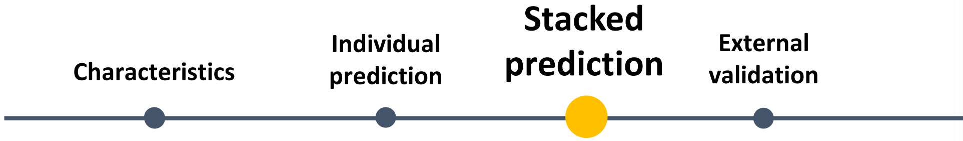


Table 4. The performance of final classifier by time at risk window

190911. revised

Graph	TAR (days)	Accuracy	Macro F1	Mean AUPRC	Mean AUROC
A	30	0.9373	0.3380	0.6440	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

Validation – PR curves


Characteristics

Individual prediction

Stacked prediction

External validation

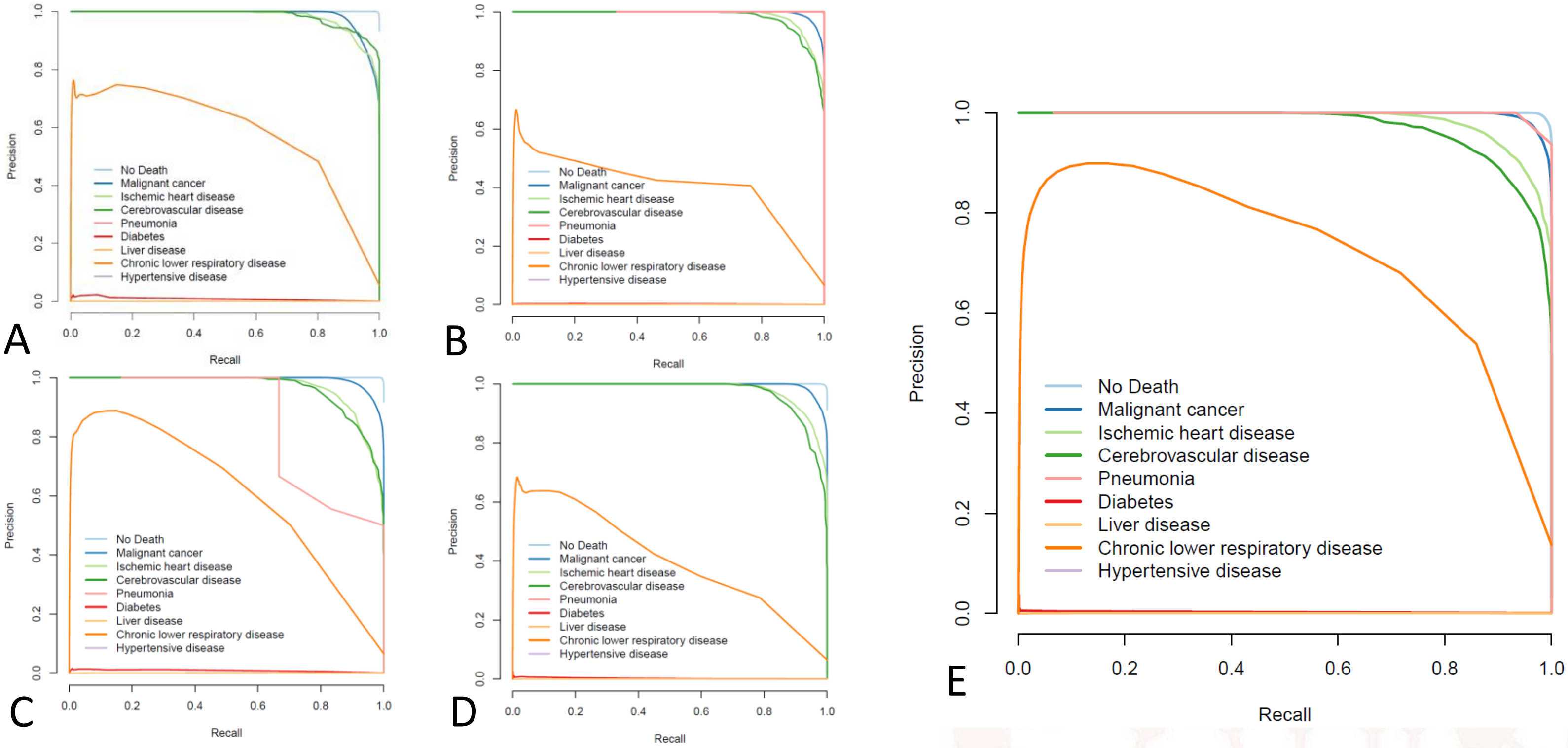


Table 4. The performance of final classifier by time at risk window 190911. revised

Graph	TAR (days)	Accuracy	Macro F1	Mean AUPRC	Mean AUROC
A	30	0.9373	0.3380	0.6440	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

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