Journal of the American Medical Informatics Association, 28(6), 2021, 1098–1107
doi: 10.1093/jamia/ocaa277
Advance Access Publication Date: 19 November 2020
Research and Applications

Research and Applications

Machine-learning model to predict the cause of death using a stacking ensemble method for observational data

Chungsoo Kim [1,†] Seng Chan You,[2,†] Jenna M. Reps [3] Jae Youn Cheong,[4] and Rae Woong Park[1,2]

[1]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Gyeonggi-do, Republic of Korea, [2]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Gyeonggi-do, Republic of Korea, [3]Janssen Research and Development, Titusville, NJ, USA, and [4]Department of Gastroenterology, Ajou University School of Medicine, Suwon, Gyeonggi-do, Republic of Korea

Corresponding Author: Rae Woong Park, MD, PhD, Department of Biomedical Informatics, Ajou University School of Medicine, 206, World cup-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16499, Republic of Korea (veritas@ajou.ac.kr)
[†]The first 2 authors contributed equally.

Received 15 June 2020; Editorial Decision 21 October 2020; Accepted 23 October 2020

ABSTRACT

Objective: Cause of death is used as an important outcome of clinical research; however, access to cause-of-death data is limited. This study aimed to develop and validate a machine-learning model that predicts the cause of death from the patient's last medical checkup.

Materials and Methods: To classify the mortality status and each individual cause of death, we used a stacking ensemble method. The prediction outcomes were all-cause mortality, 8 leading causes of death in South Korea, and other causes. The clinical data of study populations were extracted from the national claims (n = 174 747) and electronic health records (n = 729 065) and were used for model development and external validation. Moreover, we imputed the cause of death from the data of 3 US claims databases (n = 994 518, 995 372, and 407 604, respectively). All databases were formatted to the Observational Medical Outcomes Partnership Common Data Model.

Results: The generalized area under the receiver operating characteristic curve (AUROC) of the model predicting the cause of death within 60 days was 0.9511. Moreover, the AUROC of the external validation was 0.8887. Among the causes of death imputed in the Medicare Supplemental database, 11.32% of deaths were due to malignant neoplastic disease.

Discussion: This study showed the potential of machine-learning models as a new alternative to address the lack of access to cause-of-death data. All processes were disclosed to maintain transparency, and the model was easily applicable to other institutions.

Conclusion: A machine-learning model with competent performance was developed to predict cause of death.

Key words: cause of death, mortality, machine learning, classification, decision support systems, clinical

INTRODUCTION

Mortality is one of the most important end points in clinical studies aimed at determining the severity of a disease and the effectiveness of medical interventions, considering that it can be identified clearly without bias as an ultimate goal of the healthcare service.[1,2] However, all-cause mortality might not be sufficiently sensitive to iden-

tify the true effect of specific medical interventions.[3] Hence, in many clinical trials or observational studies, cause-specific mortality is a better option as a primary outcome than all-cause mortality.[4–6] Moreover, the cause-specific mortality has been a better indicator to identify disease burdens and determine the direction of health compared with all-cause mortality.[7–9]

Despite its importance, the use of cause-of-death data in observational studies has several corresponding challenges. Access to mortality data is often limited because of concerns about the exploitation of personal information.[10] Furthermore, the cause of death cannot be ascertained in most cases, even if researchers could obtain information about the mortality status of study subjects.[11,12] Notwithstanding the poor supply of mortality and cause-of-death data, various attempts have been made to overcome this obstacle and use these data in research. For example, several observational studies have been performed to link multiple data sources by national agencies[13–15] and to develop rule-based identification algorithms to pinpoint specific causes of death.[16–18]

Machine learning is widely used for the development of predictive models using large medical data sets; it has also been used in an attempt to predict a patient's mortality status. The performance of the predictive models was moderate and mainly limited to the presence of death within certain conditions, especially in-hospital death.[19–22] In addition, most of the developed machine-learning models are not spreading in clinical settings, even though they exhibit impressive performance, because of their limited reproducibility and applicability.[23] Reps et al developed a model that predicts whether the end of observation is caused by the patient's death or loss of observation by employing the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). That study was limited by the fact that the predicted outcome was only the presence of death; nevertheless, its performance was highly discriminative and the study was fully reproducible.[20] Using the OMOP-CDM not only facilitates the development and validation of models by standardizing the structure and meaning of data but also reduces the probability of the errors that occur during replication studies.

Although various studies are currently underway, a machine-learning model that can predict a patient's cause of death with sufficient transparency and applicability has yet to be developed. Hence, our study aimed to develop and validate a model for predicting the cause of death that leverages machine-learning techniques and evaluate the feasibility of the developed model on data without a known cause of death by inspecting data imputation.

MATERIALS AND METHODS

We employed the OMOP-CDM and Patient-Level Prediction (PLP) frameworks offered by Observational Health Data Sciences and Informatics (OHDSI) to develop and validate predictive models. The PLP framework consists of standardized model development and validation processes that require defining predictable problems and selecting the target population, outcome, population settings, predictors, and statistical algorithms.[24,25] Considering that the current PLP framework can only predict a binary outcome, we developed an ensemble method to predict multiclass outcomes. We developed the prediction model using the claims database of South Korea, then validated the model using the electronic health record (EHR) database of a tertiary teaching hospital. Subsequently, an imputation process was performed for 3 US claims databases that have no cause-of-death data.

All models were developed by using the R software version 3.4.4 (R Foundation for Statistical Computing, Vienna, Austria). We shared all codes used to develop the prediction model and the whole package via an online source code repository (https://github.com/ABMI/CauseSpecificMortality), and the readily executable model and computational environment for reproducibility were released by Docker (Docker image: ted9219/causespecificmortality).[26] The institutional review board at Ajou University Hospital of the Republic of Korea approved this study (IRB approval number: AJIRB-MED-MDB-19-527).

Data sources

The South Korean National Health Insurance System–National Sample Cohort (NHIS–NSC) database includes the sampled claims data of 2.2% of the total eligible Korean population in 2002.[27,28] It contains follow-up data from 1 125 691 patients recorded from 2002 to 2013. The cause-of-death data were collected from the cause-of-death database of Statistics Korea, which is linked to the NHIS–NSC. The NHIS–NSC database was converted into OMOP-CDM version 5.3.[29]

The Ajou University School of Medicine (AUSOM) database is the EHR database of 2 940 379 patients who visited the Ajou University Medical Center from 1994 to 2017. The cause-of-death records were ascertained from the death certificates of AUSOM issued by attending physicians. The AUSOM database is also in the form of the OMOP-CDM version 5.3.

Optum's De-Identified Clinformatics Data Mart Database–Date of Death (Optum DOD) is a US administrative claims database that includes over 83 million patient records collected from 2000 to 2019. The Optum DOD table has death records sourced from the Death Master File maintained by the Social Security Office of the US. The database has complete death records up to 2013 and partial death records after 2013 but does not contain any cause-of-death data.

The IBM MarketScan Medicare Supplemental Database (MDCR) represents the health services of retirees in the US with primary or Medicare supplemental coverage. The database contains the records of 10 088 000 patients collected between 2000 and 2019. The IBM MarketScan Multi-State Medicaid Database (MDCD) contains US health insurance claims for Medicaid enrollees from multiple states and includes 28 777 000 individuals with data recorded between 2006 and 2019. These 2 IBM databases include death records at discharge (in-hospital death only) and no cause of death data. The summaries of all of these databases are presented in Table 1.

Target population

We identified patients with health records spanning more than 1 year in the NHIS–NSC database as our target population. We defined the date of patients' last visit to their healthcare provider as the index date. We excluded patients who had 1 year or less of observation prior to the index date, for securing sufficient data. Moreover, any patient with an index date within 1 year of the end of the database was excluded to avoid bias from right censoring. Hence, patients whose last healthcare claims in the NHIS–NSC database were recorded in 2012 or before were included in the target population (Figure 1).

Table 1. Summary of databases for model development, validation, and imputation

Data source	Data type	No. of patients, n	No. of target population, n	Time, year	
				Start	End
NHIS-NSC	Claims	1 125 691	174 747	2002	2013
AUSOM	EHR	2 940 379	729 065	1994	2017
MDCD	Claims	28 777 000	995 372	2006	2019
MDCR	Claims	10 088 000	994 518	2000	2019
Optum DOD	Claims	83 650 000	407 604	2000	2019

Abbreviations: AUSOM, Ajou University School Of Medicine; EHR, electronic health record; MDCD, IBM MarketScan Multi-State Medicaid Database; MDCR, IBM Market Scan Medicare Supplemental Database; NHIS-NSC, National Health Insurance Services-National Sample Cohort.

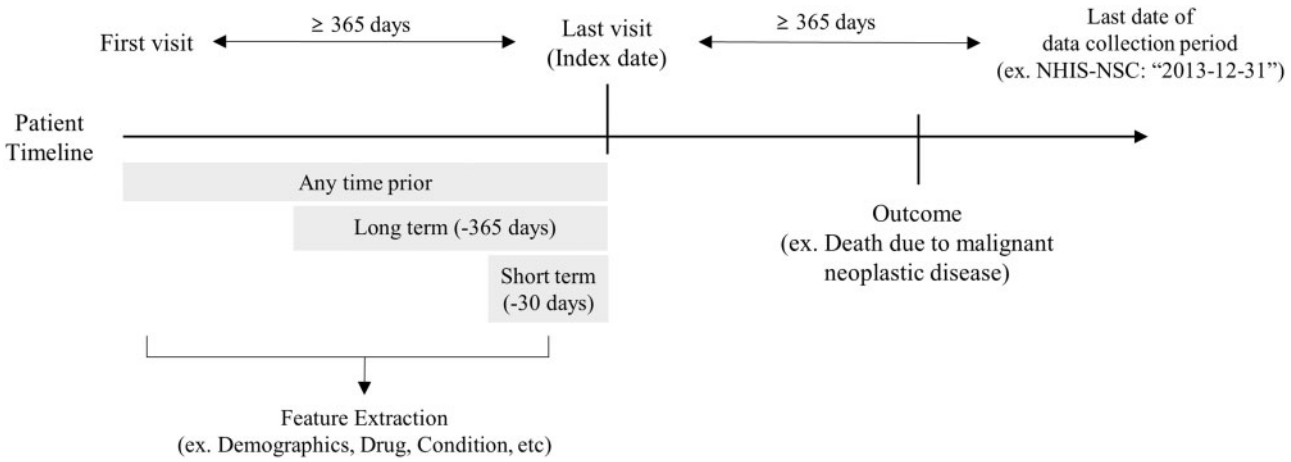

Figure 1. Target population criteria and feature extraction for base learner development. The patient's index date was set as the date of the last visit to healthcare provider, and the patients with intervals ≥ 1 year from the first visit were extracted. In addition, patients who visited during the last year of the database were excluded from the target population to prevent the bias due to censoring of the records. The outcome was determined to have "occurred" when within a certain time-at-risk interval after the index date. The feature of the patients was collected before the index date, and features within the long-term and short-term prior index date were also collected for the temporality.

Abbreviation: NHIS–NSC, National Health Insurance System–National Sample Cohort.

Outcome

We defined 10 outcomes, including mortality *per se* and South Korea's 8 leading causes of death (ie, malignant neoplastic disease, cerebrovascular disease, ischemic heart disease, pneumonia, chronic lower respiratory disease, liver disease, diabetes mellitus, hypertensive disease) and other causes.[30] Deaths not included in the 8 leading causes of death were considered "other causes". Supplementary Table 1 lists the International Classification of Disease (ICD) 10th revision code sets for each cause of death. We used the Systematized Nomenclature of Medicine Clinical Term (SNOMED-CT) equivalent of the cause of death ICD-10 code set for the population extraction.[31] Patients were recognized as "dead" if they had a death record within the specific period after the index date (time-at-risk). We employed various time-at-risk periods (30, 60, 90, 180, and 365 days) and set 60 days as the primary time-at-risk period.

Predictors

Patient demographics (gender, age, and age in 5-year groups), condition (medical diagnosis), condition group (grouped using a SNOMED-CT hierarchy), drug, drug group (grouped into ingredients), measurement, procedure, observation (eg, questionnaire answers or income status), device, and visit count were used as the input features of models. The constructed indicator features considered a missing condition record as the absence of the condition. With the exception of the demographics, all features were extracted on the basis of not only records at any time prior (all days before) but also records in the specific periods of long-term (−365 days) prior or short-term (−30 days) prior to the index date. This extraction strategy aimed to capture the temporality of the patient's medical history. Furthermore, we developed a model without temporal features under the same condition and compared the performances of the models to confirm the effect of the temporal features on such performances. We also calculated the relative importance of variables. Importance was calculated by beta coefficient for logistic regression or information gain for boosting algorithms.

Model development

For model development, we used the stacking ensemble (or stacked generalization) method.[32] The stacking ensemble method is a type of model ensemble method, using predicted probabilities from the individual classifiers (base learners) as trainable features for meta-learners. The performance of stacking ensemble methods was shown to be more robust than the individual classifier in several prior studies.[33–35] Thus, we developed the 2-level stacking ensemble model

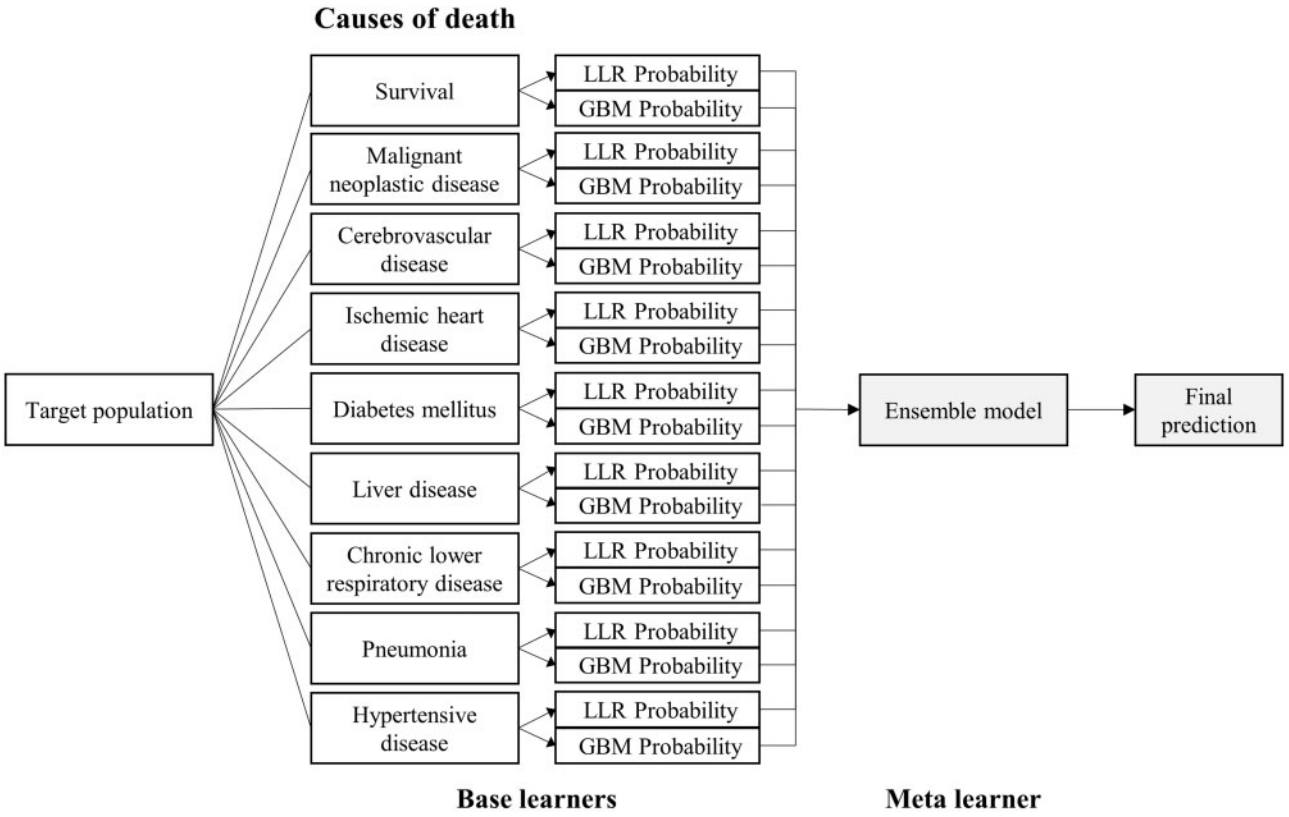

Figure 2. The schematic view of the stacking ensemble model architecture. A **2**-level stacking ensemble method was used to predict the patient's cause of death. The stacking model consists of base learners and meta-learner, and the meta-learner uses the prediction results of the base learners as input variables. Base learners that predict each of the survival and **8** causes of death as an outcome of prediction are developed first by applying **2** algorithms, lasso logistic regression and gradient boosting machine. For meta-learner, 18 input variables from base learners are used to make the final prediction.

Abbreviations: GBM, gradient boosting machine; LLR, lasso logistic regression.

consisting of base learners and a single meta-learner. For the base learners, we employed 2 machine-learning techniques (ie, the lasso logistic regression [LLR] and the gradient boosting machine [GBM]). Using either LLR or GBM, each base learner produces an estimate for a given binary outcome in the survival of patients (mortality *per se*) or 8 causes of death. If a patient was predicted to be dead by a base learner, and the cause was not classified into 1 of 8 specified causes, the model classified the cause of death for a patient as "other causes" (Figure 2). We split the data set into training and test sets by using a 75:25 proportion and then performed 3-fold cross-validation.

Performance evaluation

We calculated 4 metrics according to the one-versus-all and Hand and Till (2001) approach using the following multiple classification models: accuracy, F1 score, the mean area under the precision and recall curve (mean AUPRC), and the generalized area under the receiver operating characteristic curve (AUROC).[36,37] The accuracy is defined as the proportion with correctly predicted actual statuses. The F1 score is the harmonic mean of precision and recall. $AUPRC_i$ was calculated for each cause of death and then the mean of these $AUPRC_i$s was calculated as the AUPRC. The generalized AUROC was calculated for the multiclass classification.[37] We determined the predicted cause of death as the cause having the highest probability predicted by the model, and the performance metrics were calculated on the basis of these results. To identify the optimal algorithm of the meta-learner, we compared the performance of 3 different machine-learning techniques (random forest [RF], GBM, and extreme gradient boosting [Xgboost]). We considered the AUROC as the primary criterion for selecting an algorithm.

External validation and cause of death imputation

To confirm the transportability to, and validity of the model performance in, an EHR database rather than a claims database, we conducted external validation using the AUSOM database. All settings and evaluation processes were carried out in the manner employed in the development stage. In order to prevent misassessment of performance, the cause of death was excluded from the development and evaluation when there were < 20 specific outcomes to be predicted.

Our model was used to impute the cause of death across 3 US claims databases (MDCD, MDCR, and Optum DOD) which contained no cause-of-death data. The frequency and distribution of imputed causes according to year and age group were investigated because there was no label for evaluating the model performance. Validation and imputation packages were shared via repositories (github.com/ABMI/validationCauseSpecificMortality; github.com/ABMI/CauseOfDeathImputation)

RESULTS

Population demographics

Of the 1 125 691 patients recorded in the NHIS–NSC, 1 091 418 had medical records spanning more than 1 year. Among them,

174 747 patients were selected as the target population according to the criteria described above. The number of patients who died within a given time-at-risk, ie, 30, 60, 90, 180, and 365 days, was 30 878, 35 708, 37 040, 38 862, and 40 649, respectively (Supplementary Table 2).

Within 60 days after the index date, 11 527 (32.3%), 4057 (11.4%), and 2012 (5.6%) of deaths were caused by malignant neoplastic disease, cerebrovascular diseases, and ischemic heart disease, respectively. Deaths were also caused by diabetes mellitus (n = 1631, 4.6%), liver diseases (n = 1177, 3.3%), chronic lower respiratory diseases (n = 1102, 3.1%), pneumonia (n = 880, 2.5%), and hypertensive diseases (n = 721, 2.0%) (Table 2). During 2012, 4751 deaths were recorded, which was the highest value among all follow-up years. Regarding the distribution of death according to age group, the number of deaths in the 70s was the highest, at 10 489. Malignant neoplastic disease was the most frequent cause of death in all years or in all age groups (Table 2). Of the total deaths, 43.6% were that of females. Deaths from malignant neoplastic disease accounted for 37.0% of all deaths in men and 26.2% in women. (Table 2).

Model performance

The model had an accuracy of 0.9402, an F1 score of 0.6918, a mean AUPRC of 0.9947, and an AUROC of 0.9511 (Table 3). Figure 3 depicts the ROC curve of the final stacking ensemble model which was obtained using the Xgboost algorithm. Supplementary Table 3 summarizes the overall results of the performance of the stacking ensemble models according to the meta-learner algorithm and the time-at-risk period. Regardless of algorithm, the F1 score, mean AUPRC, and AUROC were highest in the model that predicted the cause of death within 60 days. Supplementary Figure 1 depicts the ROC curves of the final model according to the different time-at-risk periods. The result of comparison between the stacking ensemble method and individual multiclass classifier are shown in the Supplementary Figure 2.

Model specification

We employed 20 719 predictors to develop the base learners. Supplementary Table 4 lists the top 10 covariates used by all base learners ranked by relative importance. Diagnosis of causative disease (same disease condition as cause-of-death such as "heart failure" for heart disease death), age in years, and diagnosis of malignant neoplastic disease with temporality were the covariates of high relative importance in all base learners. There were no large differences between the presence and absence of covariates directly related to death (Supplementary Table 5). Excluding the base learners that predicted survival and death caused by malignant neoplastic disease, the presence of a cancer diagnosis exhibited a negative association with all LLR base learners. For the meta-learner, estimates calculated from the base learners that predicted survival and malignant neoplastic disease had a high relative importance (Supplementary Table 6).

The models that did not employ the temporal features all exhibited a lower performance (Accuracy of 0.9300, F1 score of 0.6109, mean AUPRC of 0.9856, and AUROC of 0.9293) than for those that did (Supplementary Table 7).

External validation

External validation was performed by applying the stacking ensemble model to the AUSOM database. In the AUSOM database, the total size of the target population was 729 065 individuals. Among these patients, 9917 died within 60 days from the index date. The most common causes of death were malignant neoplastic disease (n = 2712, 27.4%), pneumonia (n = 1006, 10.1%), and liver disease (n = 453, 4.6%) (Supplementary Table 2). No deaths were caused by diabetes mellitus or hypertensive disease in the AUSOM database; therefore, we excluded the corresponding prediction models from the validation (Supplementary Table 2). The external validation performances had an accuracy of 0.9235, an F1 score of 0.3360, a mean AUPRC of 0.6682, and an AUROC of 0.8601 under conditions predicting the cause of death within 60 days (Table 3, Figure 3).

Cause-of-death imputation

The number of patients in target cohorts from 3 US databases was 994 518 in MDCR, 995 372 MDCD, and 407 604 in Optum DOD (Table 4). In MDCR, 301 641 patients were predicted to have died within 60 days from the last visit. Of these, 34 135 (11.32%) deaths were caused by malignant neoplastic disease, 25 421 (8.43%) were caused by chronic lower respiratory disease, and 19 184 (6.36%) were caused by diabetes mellitus. In MDCD, a total of 57 055 patients were predicted to have died. Chronic lower respiratory disease (4465, 7.83%) appeared more frequently than did other causes of death. Moreover, 20 331 patients were predicted to have died in the Optum DOD database. The 2 base learners (LLR and GBM) that predicted survival derived from the Optum DOD exhibited an AUROC of 0.9884 and 0.9881, respectively. Similar to that imputed for the MDCR, malignant neoplastic disease was the leading cause of death in Optum DOD (2222, 10.93%) (Table 4). The trend of causes of death in the 3 US databases and the NHIS–NSC database according to year and age group is shown in Figure 4.

DISCUSSION

To the best of our knowledge, this study was the first to attempt to predict the individual cause of death from an observational database. We implemented machine learning to develop a stacking ensemble method to predict death and its causes. In addition, we performed external validation and imputation across data sets collected for different purposes and from different countries. The developed model showed competent performances in terms of accuracy, mean AUPRC, AUROC (all > 0.9), and F1 score (> 0.6). The external validation, which was performed using an EHR database, exhibited promising performance in AUROC (> 0.8), but the F1 score was lower than expected. Moreover, imputation using US databases could actually be performed. The fact that our model could be applied to databases other than the development database indicates that our model is transportable and applicable throughout the OMOP-CDM and PLP frameworks. Taken together, our findings suggest that the model can be used as a predictive tool to resolve a common limitation of observational studies (ie, the lack of cause-of-death data).[38,39]

Implications

The values of the model developed in this study are its high scalability and transparent development process. We intended to use the stacking ensemble method for the multiclass classification problem and applied it to the OHDSI PLP framework. Considering that the model is composed of binary prediction models, it can be easily extended by changing or adding the base learners representing each

Table 2. Temporal and demographic group trend of cause-of-death data in target cohort from NHIS–NSC

Cause of death	Total (n, %)	NHIS-NSC (n = 174 747)																					
		By year (n, %)										By age groups (n, %)										By sex (n, %)	
		2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	Female	Male
Total death	35 708	2278	2930	3122	3300	3462	3702	3786	3873	4504	4751	97	137	357	815	2078	3698	6494	10 489	9642	1901	15 576	20 132
Cancer	11 527	818	1012	1007	1113	1101	1245	1197	1206	1379	1449	21	28	65	227	739	1658	2993	3666	1948	182	4078 (26.2)	7449 (37.0)
	(32.3)	(35.9)	(34.5)	(32.3)	(33.7)	(31.8)	(33.6)	(31.6)	(31.1)	(30.6)	(30.5)	(21.7)	(20.4)	(18.2)	(27.9)	(35.6)	(44.8)	(46.1)	(35.0)	(20.2)	(9.6)		
CeVD	4057	285	355	411	395	414	414	403	432	484	464	0	1	5	37	156	260	661	1438	1330	169	2106 (13.5)	1951 (9.7)
	(11.4)	(12.5)	(12.1)	(13.2)	(12.0)	(12.0)	(11.2)	(10.6)	(11.2)	(10.8)	(9.8)	(0.0)	(0.7)	(1.4)	(4.5)	(7.5)	(7.0)	(10.2)	(13.7)	(13.7)	(8.9)		
IHD	2012	104	155	190	210	226	204	230	204	235	254	0	0	2	20	80	163	320	639	682	106	910 (5.8)	1102 (5.5)
	(5.6)	(4.6)	(5.3)	(6.1)	(6.4)	(6.5)	(5.5)	(6.1)	(5.3)	(5.2)	(5.4)	(0.0)	(0.0)	(0.6)	(2.5)	(3.9)	(4.4)	(4.9)	(6.1)	(7.1)	(5.6)		
DM	1631	130	152	157	176	153	150	167	154	161	231	0	2	0	21	55	128	339	580	452	54	812 (5.2)	819 (4.1)
	(4.6)	(5.7)	(5.2)	(5.0)	(5.3)	(4.4)	(4.1)	(4.4)	(4.0)	(3.6)	(4.9)	(0.0)	(1.5)	(0.0)	(2.6)	(2.7)	(3.5)	(5.2)	(5.5)	(4.7)	(2.8)		
LD	1178	102	123	106	114	93	121	121	111	166	121	0	2	4	45	243	307	277	190	92	18	266 (1.7)	912 (4.5)
	(3.3)	(4.6)	(4.2)	(3.4)	(3.5)	(2.7)	(3.3)	(3.2)	(2.9)	(3.7)	(2.6)	(0.0)	(1.5)	(1.1)	(5.5)	(11.7)	(8.3)	(4.3)	(1.8)	(1.0)	(1.0)		
CLRD	1102	84	111	88	84	101	121	102	105	139	167	2	1	1	1	8	29	112	384	470	94	401 (2.6)	701 (3.5)
	(3.1)	(3.7)	(3.8)	(2.8)	(2.6)	(2.9)	(3.3)	(2.7)	(2.7)	(3.1)	(3.5)	(2.1)	(0.7)	(0.3)	(0.1)	(0.4)	(0.8)	(1.7)	(3.7)	(4.9)	(4.9)		
PNA	880	16	43	57	54	67	83	76	98	186	200	1	1	3	4	5	16	60	247	408	135	421 (2.7)	459 (2.3)
	(2.5)	(0.7)	(1.5)	(1.8)	(1.6)	(1.9)	(2.2)	(2.0)	(2.5)	(4.1)	(4.2)	(1.0)	(0.7)	(0.8)	(0.5)	(0.2)	(0.4)	(0.9)	(2.4)	(4.2)	(7.1)		
HT	721	48	63	48	55	78	78	73	83	96	99	0	0	0	6	10	16	58	197	332	102	462 (3.0)	259 (1.3)
	(2.0)	(2.1)	(2.2)	(1.5)	(1.7)	(2.3)	(2.1)	(1.9)	(2.1)	(2.1)	(2.1)	(0.0)	(0.0)	(0.0)	(0.7)	(0.5)	(0.4)	(0.9)	(1.9)	(3.4)	(5.4)		
Others	12 600	691	916	1058	1099	1229	1286	1417	1480	1658	1766	73	102	277	454	782	1121	1674	3148	3928	1041	6120 (39.3)	6480 (32.2)
	(35.3)	(30.3)	(31.3)	(33.9)	(33.3)	(35.5)	(34.7)	(37.4)	(38.2)	(36.8)	(37.2)	(75.3)	(74.5)	(77.6)	(55.7)	(37.6)	(30.3)	(25.8)	(30.0)	(40.7)	(54.8)		

Abbreviations: Cancer, malignant neoplastic disease; CeVD, Cerebrovascular disease; CLRD, Chronic lower respiratory disease; DM, diabetes mellitus; HT, hypertensive disease; IHD, ischemic heart disease; LD, liver disease; NHIS–NSC, National Health Insurance Services–National Sample Cohort; PNA, Pneumonia.

class of the desired predictive outcome (in this study, cause-of-death). This approach may be applicable in future research to further classify the detailed causes of death. The use of OMOP-CDM enables transportable model development based on its standardized data structure and interoperability with other databases. We have disclosed the model in an online repository, to ensure transparency, and it can be applied directly by other institutions using the OMOP-CDM; therefore, our model has portability and reproducibility. Most machine-learning models developed in the medical field lack adequate reporting regarding model development or performance, thereby being of limited usefulness in practice.[23,40,41] In the current study, we attempted to overcome these issues by providing the complete details of model development via standardized data, terminology, model development framework, code, and computational environment disclosure.

Table 3. Performance results of the final ensemble model in internal and external validation databases

Performance metrics	NHIS-NSC	AUSOM
ACC	0.9402	0.9190
F1 score	0.6918	0.3131
Mean AUPRC	0.9947	0.6635
AUROC	0.9511	0.8887

Abbreviations: ACC, accuracy; AUPRC, area under the precision recall curve; AUROC, area under the receiver operating characteristics curve; AUSOM, Ajou University School Of Medicine; NHIS–NSC, National Health Insurance Services–National Sample Cohort.

Interpretation

Using the important covariates of the model, we could understand indirectly how the ensemble model predicted and classified the cause of death. In most base learners, age, causative-disease-related features, and diagnosis of malignant neoplastic disease were relatively important covariates. Age is a known predictor of any death, and the causative disease-related features could be explained sufficiently.

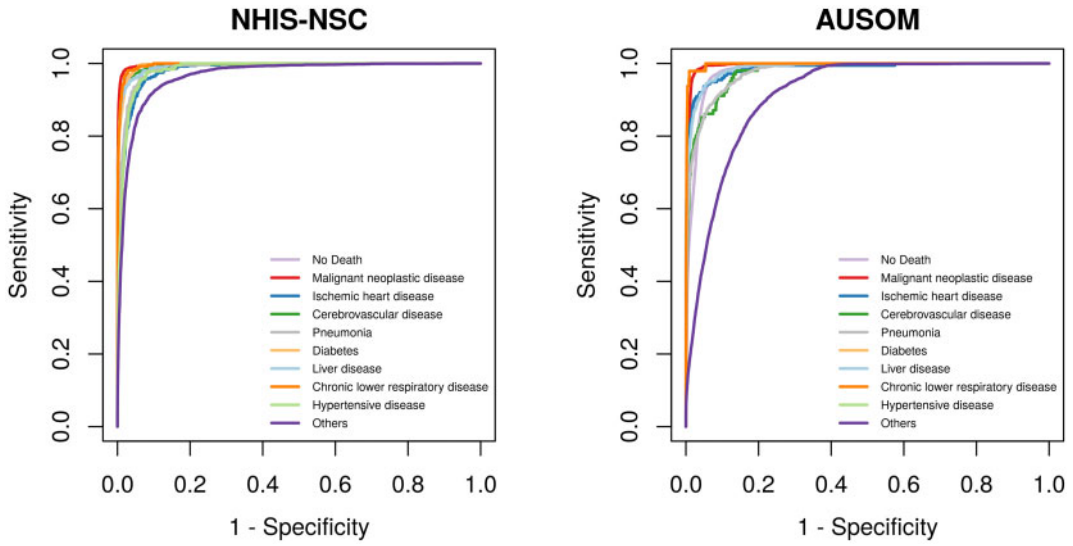

Figure 3. Receiver operating characteristic curve of the final model from development and validation datasets. The receiver operating characteristic (ROC) curve plotted from the cause of death prediction model. The presence of death within 60 days from the last visit date and its cause were predicted. As a meta-learner, Xgboost was used. The ROC curve for each cause of death is shown. The figure shows for the NHIS–NSC's test set and AUSOM dataset.

Abbreviations: AUSOM, Ajou University School of Medicine; NHIS–NSC, National Health Insurance Services–National Sample Cohort.

Table 4. Results of cause-of-death imputation for US databases having no cause-of-death data

Causes of death	Number of predicted (percent of total deaths, %)		
	MDCR (n = 994 518)	MDCD (n = 995 372)	Optum DOD (n = 407 604)
Total death	301 641	57 055	20 331
Malignant neoplastic disease	34 135 (11.32)	4378 (7.67)	2222 (10.93)
Chronic lower respiratory disease	25 421 (8.43)	4465 (7.83)	1426 (7.01)
Diabetes	19 184 (6.36)	2622 (4.60)	1139 (5.60)
Ischemic heart disease	5961 (1.98)	702 (1.23)	372 (1.83)
Cerebrovascular disease	5463 (1.81)	1484 (2.60)	417 (2.05)
Hypertensive disease	4105 (1.36)	133 (0.23)	286 (1.41)
Pneumonia	523 (0.17)	50 (0.09)	26 (0.13)
Liver disease	246 (0.08)	307 (0.54)	65 (0.32)
Other causes	206 603 (68.49)	42 914 (75.22)	14 378 (70.72)

Abbreviations: MDCD, IBM MarketScan Multi-State Medicaid Database; MDCR, IBM Market Scan Medicare Supplemental Database.

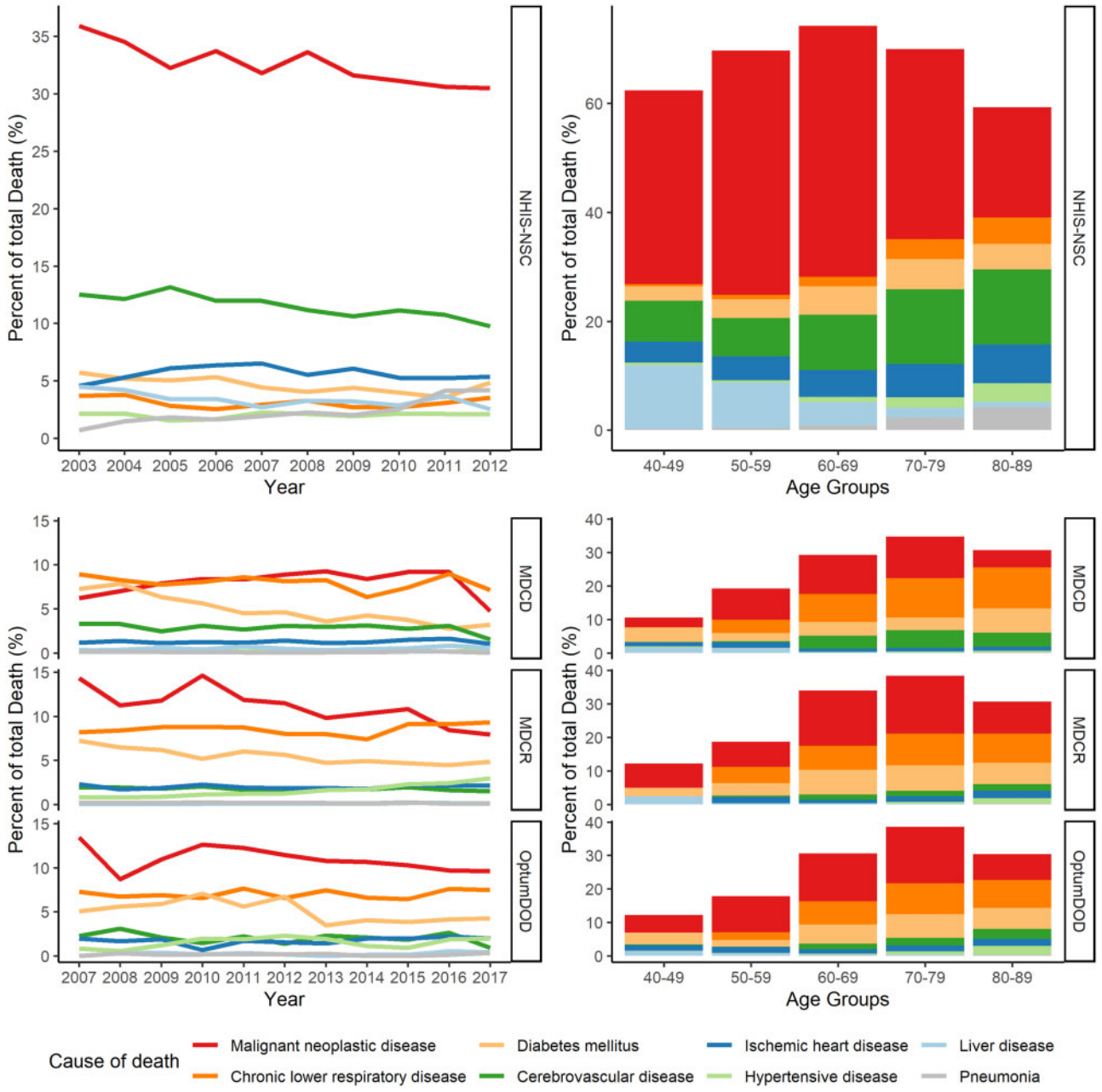

Figure 4. Cause-of-death temporal trend and demographic distribution in the NHIS–NSC and US databases, imputed by the prediction model. Distribution of causes of death according to age group and year. The graph at the top shows that malignant cancer death accounted for the largest proportion, and that this trend was independent of year and age group in the NHIS–NSC. The graph at the bottom shows the distribution of the cause of death imputed from US databases using the developed model. Because the year of each database is different, the graph is limited to the specific year and age group.

For example, subarachnoid hemorrhage diagnosis and therapeutic agents, such as mannitol, are strongly associated with death caused by cerebrovascular disease. In addition, a diagnosis of malignant neoplastic disease can be an important factor in distinguishing the cause of death from comorbidities. In fact, with the exception of the base learners of malignant neoplastic disease death, the diagnosis of malignant neoplastic disease was negatively associated with each outcome in the LLR base learners. Thus, excluding death caused by malignant neoplastic disease, which is the most common cause of death in South Korea, is important in determining a cause of death other than malignant neoplastic disease within the concept of the competitive covariates.[42] Moreover, features with temporality contribute to the discrimination between the underlying diseases and the cause of death, thus enabling a more accurate prediction.[43] Overall, the model was sufficiently interpretable and clinically valid.

We performed an external validation process with the developed model. The performance results showed an encouraging AUROC but a lower F1 score. This might be a threshold problem with predicted probabilities. More fundamentally, it is due to differences in population characteristics in the NHIS–NSC and AUSOM databases. The NHIS–NSC database reflects the nationwide mortality rate and proportion, whereas the AUSOM database has the mortality data of a single tertiary hospital. The other causes of death for NHIS–NSC and AUSOM differ by approximately 20%.

We performed a cause-of-death data imputation from US claims databases using the developed prediction model. US claims databases do not commonly have cause of death recorded. This process was used for testing the feasibility of our model to different settings in the situation that there are no ground truth labels of the cause of death. The AUROC for our model predicting mortality itself was above 0.98 in the US databases with mortality information. In the NHIS–NSC database, the 3 leading causes of death were malignant neoplastic disease, cerebrovascular disease, and ischemic heart disease. These results were in line with reports from Statistics Korea.[30] In the US databases, malignant neoplastic disease, chronic lower respiratory disease, and diabetes mellitus were the most common causes of death predicted by our model. This finding differs from the well-known ranking of causes of mortality in the US: 1. heart disease, 2. cancer, and 3. chronic lower respiratory disease.[44] This discrepancy might be attributable to the difference in the definition of disease between Statistics Korea and the US National Center for Health Statistics. The definition of heart disease death in the US reports includes death due to disease coded by I00-I09, I11, I13, and I20-I51 (ICD-10), whereas the Korean definition includes only ICD-10 codes I20-25 and I30-52.[44] In our sensitivity studies, the imputation using a model with a modified heart disease definition according to the US definition resulted in an increase in the rate of heart disease death in all databases, and the MDCR database showed the most similar results to the US statistics (Supplementary Figure 4). Additionally, the results from the model developed for predicting the cause of death only (not including predicting mortality) in the death population seemed more compatible with US statistics (Supplementary Table 8, Supplementary Figure 5).

These results and interpretations suggest that our model is applicable to other databases and other countries, and that the imputed results might be reliable.

Limitations

Regarding the study's limitations, this model depended on the quality of the cause-of-death data. Misclassification of the cause of death may affect model performance. However, in this study, our model was developed using long-term nationwide observational cohort data with almost no missing details in the cause-of-death data. The cause-of-death data provided by the national statistics office were utilized, and the composition of the cause of death in the NHIS–NSC is similar to the cause-of-death statistics report published by Statistics Korea. Moreover, since it is difficult to accurately identify the cause of death, physicians generally make decisions based on health records except for autopsy cases. Similarly, the machine-learning models applied in this study also judged using specific covariates they decided as important from the patient's entire health record.

Paradoxically, the lack of access to cause-of-death data implies the absence of data that can be used to evaluate and validate the performance of the model, which itself represents a limitation of the research. And each database we validated and imputed cause-of-death data has different characteristics of the patient group; therefore, the limitation remains that our results cannot be accurately matched with the cause-of-death statistics. Further external validation using data from additional countries and institutions is required to increase the generalizability of our model, even though the model has undergone a process of external validation and imputation across databases.

CONCLUSION

A machine-learning model for predicting cause of death was developed using a standardized common data model and an extensible analytical method. We attempted to use a transparent development process; consequently, the prediction performance of the model was impressive. The majority of observational data sets lack cause-of-death data and our model can be used to impute this information with high discriminative performance.

FUNDING

This work was supported by the Bio Industrial Strategic Technology Development Program (20001234, 20003883) 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].

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and design of the work, the analysis, and the interpretation of data for the work. All authors contributed in drafting, revising, and approving the final version.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

This study used NHIS-NSC data generated by the National Health Insurance Service (NHIS-2020-2-123).

CONFLICT OF INTEREST STATEMENT

Jenna M Reps is an employee of Janssen Research & Development and shareholder of Johnson & Johnson. The other authors have no conflict of interest.

REFERENCES

1. Weiss NS. All-cause mortality as an outcome in epidemiologic studies: proceed with caution. *Eur J Epidemiol* 2014; 29 (3): 147–9.
2. Black WC, Haggstrom DA, Welch HG. All-cause mortality in randomized trials of cancer screening. *J Natl Cancer Inst* 2002; 94 (3): 167–73.
3. Sasieni PD, Wald NJ. Should a reduction in all-cause mortality be the goal when assessing preventive medical therapies? *Circulation* 2017; 135 (21): 1985–7.
4. Heijnsdijk EAM, Csanádi M, Gini A, *et al.* All-cause mortality versus cancer-specific mortality as outcome in cancer screening trials: a review and modeling study. *Cancer Med* 2019; 8 (13): 6127–38.
5. Lin JS, Piper MA, Perdue LA, *et al.* Screening for colorectal cancer: updated evidence report and systematic review for the US preventive services task force. *JAMA* 2016; 315 (23): 2576–94.
6. Yusuf S, Negassa A. Choice of clinical outcomes in randomized trials of heart failure therapies: disease-specific or overall outcomes? *Am Heart J* 2002; 143 (1): 22–8.
7. Roth GA, Abate D, Abate KH, *et al.* Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 2018; 392 (10159): 1736–88.
8. Starfield B. Is US health really the best in the world? *JAMA* 2000; 284 (4): 483–5.

9. Murray CJ, Lopez AD. Evidence-based health policy–lessons from the Global Burden of Disease Study. *Science* 1996; 274 (5288): 740–3.
10. Levin MA, Lin H-M, Prabhakar G, *et al.* Alive or dead: validity of the social security administration death master file after 2011. *Health Serv Res* 2019; 54 (1): 24–33.
11. Ooba N, Setoguchi S, Ando T, *et al.* Claims-based definition of death in Japanese claims database: validity and implications. *PLoS One* 2013; 8 (5): e66116.
12. Bhalla K, Harrison JE, Shahraz S, *et al.* Availability and quality of cause-of-death data for estimating the global burden of injuries. *Bull World Health Organ* 2010; 88 (11): 831–8.
13. Lin L-y, Warren-Gash C, Smeeth L, *et al.* Data resource profile: the National Health Insurance Research Database (NHIRD). *Epidemiol Health* 2018; 40: e2018062-0.
14. Bezin J, Duong M, Lassalle R, *et al.* The national healthcare system claims databases in France, SNIIRAM and EGB: powerful tools for pharmacoepidemiology. *Pharmacoepidemiol Drug Saf* 2017; 26 (8): 954–62.
15. Ohlmeier C, Langner I, Garbe E, *et al.* Validating mortality in the German Pharmacoepidemiological Research Database (GePaRD) against a mortality registry. *Pharmacoepidemiol Drug Saf* 2016; 25 (7): 778–84.
16. Singh S, Fouayzi H, Anzuoni K, *et al.* Diagnostic algorithms for cardiovascular death in administrative claims databases: a systematic review. *Drug Saf* 2019; 42 (4): 515–27.
17. Langner I, Ohlmeier C, Haug U, *et al.* Implementation of an algorithm for the identification of breast cancer deaths in German health insurance claims data: a validation study based on a record linkage with administrative mortality data. *BMJ Open* 2019; 9 (7): e026834.
18. Gagnon B, Mayo NE, Laurin C, *et al.* Identification in administrative databases of women dying of breast cancer. *J Clin Oncol* 2006; 24 (6): 856–62.
19. Weng SF, Vaz L, Qureshi N, *et al.* Prediction of premature all-cause mortality: a prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS One* 2019; 14 (3): e0214365.
20. Reps JM, Rijnbeek PR, Ryan PB. Identifying the DEAD: development and validation of a patient-level model to predict death status in population-level claims data. *Drug Saf* 2019; 42 (11): 1377–86.
21. Rajkomar A, Oren E, Chen K, *et al.* Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018; 1: 18.
22. Taylor RA, Pare JR, Venkatesh AK, *et al.* Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven machine learning approach. *Acad Emerg Med* 2016; 23 (3): 269–78.
23. Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. *JAMA* 2020; 323 (4): 305.
24. Reps JM, Schuemie MJ, Suchard MA, *et al.* Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc* 2018; 25 (8): 969–75.
25. Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015; 216: 574–8.
26. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. *Linux J* 2014; 2014 (239): 2.
27. Lee S, Lee YB, Kim BJ, *et al.* All-cause and cause-specific mortality risks associated with alopecia areata: a Korean nationwide population-based study. *JAMA Dermatol* 2019; 155 (8): 922–8.
28. Lee J, Lee JS, Park SH, *et al.* Cohort Profile: The National Health Insurance Service-National Sample Cohort (NHIS-NSC), South Korea. *Int J Epidemiol* 2017; 46 (2): e15.
29. You SC, Lee S, Cho S-Y, *et al.* Conversion of National Health Insurance Service-National Sample Cohort (NHIS-NSC) database into Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM). *Stud Health Technol Inform* 2017; 245: 467–70.
30. Statistics Korea. KOSIS. http://kosis.krAccessed June 5, 2020
31. Hripcsak G, Levine ME, Shang N, *et al.* Effect of vocabulary mapping for conditions on phenotype cohorts. *J Am Med Inform Assoc* 2018; 25 (12): 1618–25.
32. Wolpert DH. Stacked generalization. *Neural Netw* 1992; 5 (2): 241–59.
33. Kang S, Cho S, Kang P. Multi-class classification via heterogeneous ensemble of one-class classifiers. *Eng Appl Artific Intell* 2015; 43: 35–43.
34. Zhai B, Chen J. Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China. *Sci Total Environ* 2018; 635: 644–58.
35. Wang Y, Wang D, Geng N, Wang Y, Yin Y, Jin Y. Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. *Appl Soft Comput* 2019; 77: 188–204.
36. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 2009; 45 (4): 427–37.
37. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach Learn* 2001; 45 (2): 171–86.
38. Turgeon RD, Koshman SL, Youngson E, *et al.* Association of Ticagrelor vs Clopidogrel with major adverse coronary events in patients with acute coronary syndrome undergoing percutaneous coronary intervention. *JAMA Intern Med* 2020; 180 (3): 420.
39. Zeng C, Dubreuil M, LaRochelle MR, *et al.* Association of tramadol with all-cause mortality among patients with osteoarthritis. *JAMA* 2019; 321 (10): 969–82.
40. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *The Lancet* 2019; 393 (10181): 1577–9.
41. Collins GS, Reitsma JB, Altman DG, *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* 2015; 13 (1): 1.
42. Satagopan JM, Ben-Porat L, Berwick M, *et al.* A note on competing risks in survival data analysis. *Br J Cancer* 2004; 91 (7): 1229–35.
43. Balabaeva K, Kovalchuk S. Comparison of temporal and non-temporal features effect on machine learning models quality and interpretability for chronic heart failure patients. *Procedia Comput Sci* 2019; 156: 87–96.
44. Kochanek KD, Murphy SL, Xu JQ, Arias E. Deaths: Final data for 2017. *Natl Vital Stat Rep* 2019; 68 (9).