Published online: 2021-09-28

Development of Prediction Models for Unplanned Hospital Readmission within 30 Days Based on Common Data Model: A Feasibility Study

Borim Ryu[1,2] Sooyoung Yoo[1,*] Seok Kim[1] Jinwook Choi[2,3,*]

[1] Office of eHealth Research and Business, Seoul National University Bundang Hospital, Seongnam, South Korea
[2] Department of Biomedical Engineering, College of Medicine, Seoul National University, Seoul, South Korea
[3] Institute of Medical and Biological Engineering, Medical Research Center, Seoul National University, Seoul, South Korea

Address for correspondence Jinwook Choi, MD, PhD, Department of Biomedical Engineering, College of Medicine, Seoul National University, 28 Yongon-dong Chongro-gu, Seoul 110-799, South Korea (e-mail: jinchoi@snu.ac.kr).

Methods Inf Med 2021;60:e65–e75.

Abstract

Background Unplanned hospital readmission after discharge reflects low satisfaction and reliability in care and the possibility of potential medical accidents, and is thus indicative of the quality of patient care and the appropriateness of discharge plans.

Objectives The purpose of this study was to develop and validate prediction models for all-cause unplanned hospital readmissions within 30 days of discharge, based on a common data model (CDM), which can be applied to multiple institutions for efficient readmission management.

Methods Retrospective patient-level prediction models were developed based on clinical data of two tertiary general university hospitals converted into a CDM developed by Observational Medical Outcomes Partnership. Machine learning classification models based on the LASSO logistic regression model, decision tree, AdaBoost, random forest, and gradient boosting machine (GBM) were developed and tested by manipulating a set of CDM variables. An internal 10-fold cross-validation was performed on the target data of the model. To examine its transportability, the model was externally validated. Verification indicators helped evaluate the model performance based on the values of area under the curve (AUC).

Results Based on the time interval for outcome prediction, it was confirmed that the prediction model targeting the variables obtained within 30 days of discharge was the most efficient (AUC of 82.75). The external validation showed that the model is transferable, with the combination of various clinical covariates. Above all, the prediction model based on the GBM showed the highest AUC performance of 84.14 ± 0.015 for the Seoul National University Hospital cohort, yielding in 78.33 in external validation.

Conclusions This study showed that readmission prediction models developed using machine-learning techniques and CDM can be a useful tool to compare two hospitals in terms of patient-data features.

Keywords
- readmission
- machine learning
- database
- patient-level prediction
- common data model

* *These authors equally contributed to this work as co-corresponding authors.*

received
February 17, 2021
accepted after revision
July 5, 2021
published online
September 28, 2021

DOI https://doi.org/10.1055/s-0041-1735166.
ISSN 0026-1270.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Background and Significance

While unplanned hospital readmissions are frequent and costly, they are potentially avoidable.[1] Readmission to a hospital shortly after discharge refers to the case where a patient requires retreatment within a short period of time after receiving a particular treatment.[2] Receiving unplanned recare is based on the premise that there were qualitative problems with the first treatment. The factors behind hospital readmission could be numerous, complex, and interrelated.[3]

In 2012, the United States began imposing penalties on hospitals (1% of the hospital's base Medicare inpatient payments[2,4]) with high 30-day readmission rates for heart failure, acute myocardial infarction, and pneumonia; fines of about $280 million were imposed on 2,213 hospitals in that year. Later, additional penalties were added for chronic lung disease and coronary artery bypass transplantation, and the penalties were increased to 2 and 3%, respectively. Policymakers and medical institutions have proposed several programs to reduce readmission and improve accessibility.[5]

Several related studies have investigated patient-related risk factors for hospital readmissions.[6–14] Robert and Tamer[15] conducted a predictive study of readmission within 30 days using the LACE index (includes length of stay, acute admission status, Charlson comorbidity index [CCI], and emergency department visits in the past year) and HOSPITAL score (includes hemoglobin at discharge, discharge from oncology service, sodium level at discharge, procedure during index hospitalization, index hospitalization type, number of admissions in the past year, and length of stay) on patients discharged from the Memorial Medical Center from October 2015 to March 2016, with an area under the curve (AUC) value of 0.75 for the HOSPITAL score and an AUC value of 0.58 for the LACE index.[15] Miller et al evaluated the ability to independently predict hospital readmissions within 30 days and compared occupational capabilities with the LACE index to develop predictive tools to identify patients at high risk of unplanned readmissions.[16] Shameer et al demonstrated the potential biomarker sets for readmission probability.[17] Among 1,068 target patients, 178 were readmitted within 30 days (readmission rate 16.66%), and electronic medical record (EMR) data (including diagnostic codes, drugs, laboratory measurements, surgical procedures, and vitals) were extracted and used in multistage modeling using the Naive Bayes algorithm. As a result of their study, compared with the existing readmission prediction model, the EMR-wide prediction model achieved an AUC of 0.78, which was found to be an effective application of data-based machine learning.

Although many studies for predicting hospital readmission have been conducted in this manner, the developed prediction models are difficult to apply to other institutions because they were made based on data from a specific single institution. For example, to apply a machine-learning prediction algorithm to data from other hospitals, it is necessary to process data in the form of input values appropriate to the algorithm; in some cases, the program needs to be modified according to the data features of other institutions.

Health care data are collected and stored for many purposes, including (1) to directly facilitate research as a form of survey or registry data, or (2) to record the conduct of health care (usually called electronic health record [EHR]), or (3) to manage payments for health care such as claims data. All three are routinely used for clinical research (the latter two as secondary use data), and all three types have their unique content formatting and encoding.[18] The reuse of EHR data for research is a relatively new field and, so far, there has been a lack of awareness regarding the code setting engineering issue.[19] A common data standard can alleviate this need by omitting the extraction step and allowing standardized analytics to be executed on the data in its native environment, that is, the analytics come to the data instead of the other way around.[18] Within the last decade, several common data models (CDMs) have been collaboratively developed for clinical research data. These include the Sentinel CDM,[20] the National Patient-Centered Clinical Research Network CDM,[21] the Health Care Systems Research Network (formerly known as the HMO Research Network) Virtual Data Warehouse,[22] the Observational Medical Outcomes Partnership (OMOP) CDM,[23] and the Clinical Data Interchange Standards Consortium Study Data Tabulation Model.[24]

According to the literature,[25] OMOP CDM performs the best, ranking highest in a majority of the evaluation criteria when compared with the other CDM models for EHR-based longitudinal registries based on content coverage, integrity, flexibility, simplicity, integration, and implementability. To maintain and expand OMOP CDM, Observational Health Data Sciences and Informatics (OHDSI) is currently developing open-source tools for data quality and characterization, medical product safety surveillance, comparative effectiveness, quality of care, and patient-level predictive modeling.[25–30] Seoul National University Bundang Hospital (SNUBH) and Seoul National University Hospital (SNUH) EHRs have been transformed into OMOP CDM for longitudinal retrospective observation research. Once a database has been converted to OMOP CDM, it can be analyzed using OMOP CDM-based tools.

To the best of our knowledge, there has been no study until now that uses CDM data in the development of 30-day unplanned hospital readmission prediction models. In particular, the contribution of this study is to identify effective variables for hospital readmission within 30 days. As a multicenter study, we utilized data from two hospitals to carry out verification with each other and evaluate which hospital's data and models using specific variables are the most suitable models for the topic of readmission prediction. Notably, although many hospital readmission prediction models have been studied, it is difficult to apply the developed model directly to data from other hospitals, and there is a paucity of research exploring clinical variables for patient readmission within 30 days.

Objectives

In this study, we aimed to develop and validate a patient-level prediction model for all-cause 30-day hospital

readmission by applying machine-learning methods, with EMR-based clinical data converted to CDM. To the best of our knowledge, none of the studies using OMOP CDM data have explored patient-level prediction models for hospital readmission within 30 days. In this study, clinical variable combinations obtained during the period that can be effective in predicting readmissions within 30 days were considered. Therefore, the research hypothesis of this study is that the predictive model developed using CDM is easy to apply to other institutions and can explore clinical variables that might influence hospital readmission within 30 days. The CDM prediction models showed external validity in terms of model performance as well as transportability.

Methods

Study Data Description

We conducted an observational cohort study using CDM-converted EHR data from two tertiary general university hospitals: SNUH and SNUBH. These study sites have converted EHR data over a 15-year period of more than two million patients, including patient demographics, diagnosis, drug exposures, test orders and their results, surgeries, family histories, and past medical histories, into OMOP CDM. SNUH acts as the parent hospital for SNUBH and thus both use the same EHR system. Evidently, SNUH is considerably larger than SNUBH in terms of size. The parent hospital is located in the center of Seoul, and the child hospital is located in Gyeonggi-do, near Seoul, an hour away from the parent hospital.

Patients who visited the hospitals between January 1, 2017 and December 31, 2018 were included in the study. We excluded individuals who died during hospitalization and who visited a hospital for clinical trial. We organized study cohorts for patients living in Seoul or Gyeonggi Province in Korea. ►Figs. 1 and 2 show the study cohort design in this study.

Patients' clinical features were extracted from all diagnosis records prior to the end of the readmission interval, such as gender, age at visit, diagnosis history, medications, and some calculated indices such as CCI (Romano adaptation). All clinical events in OMOP CDM are expressed as concepts, which represent the semantic notion of each event. Concepts are coded as individual concept codes, higher-level concept codes, and groups of higher-level codes based on the level of standard terminologies. Each Standard Concept has a unique domain assignment, which defines which table they are recorded in. For example, signs, symptoms, and diagnosis concepts are in the Condition Domain and are recorded in "CONDITION_CONCEPT_ID" of the "CONDITION_OCCURRENCE CDM" table, which are mainly mapped into SNOMED CT standard terminologies.

We used the Feature Extraction package developed by OHDSI as an open-source CDM tool to create features for a cohort. Below is a list of different EHR-driven features with the descriptions of data used in this study. The names of OMOP CDM data tables are written in capital letters. Please refer to the OHDSI GitHub Web site for more details regarding OMOP CDM specifications.[31]

- *Diagnosis and medication information*: Records of a person suggesting the presence of a disease or medical condition stated as a diagnosis, a sign, or a symptom, which is observed by a clinician, are contained in the "CONDITION_OCCURRENCE" table. The data table "DRUG_EXPOSURE" in CDM database captures records about exposure to a drug ingested or otherwise introduced into the body. Drugs include prescription and over-the-counter medicines.

Individual diagnosis and medication records are extracted from "CONDITION_OCCURRENCE" and "DRUG_EXPOSURE" data tables in the CDM database. The table "CONDITION_ERA" includes data for the span of time the patient is assumed to have had a given condition. This data table contains chronological periods of "CONDITION_OCCURRENCE." In addition,

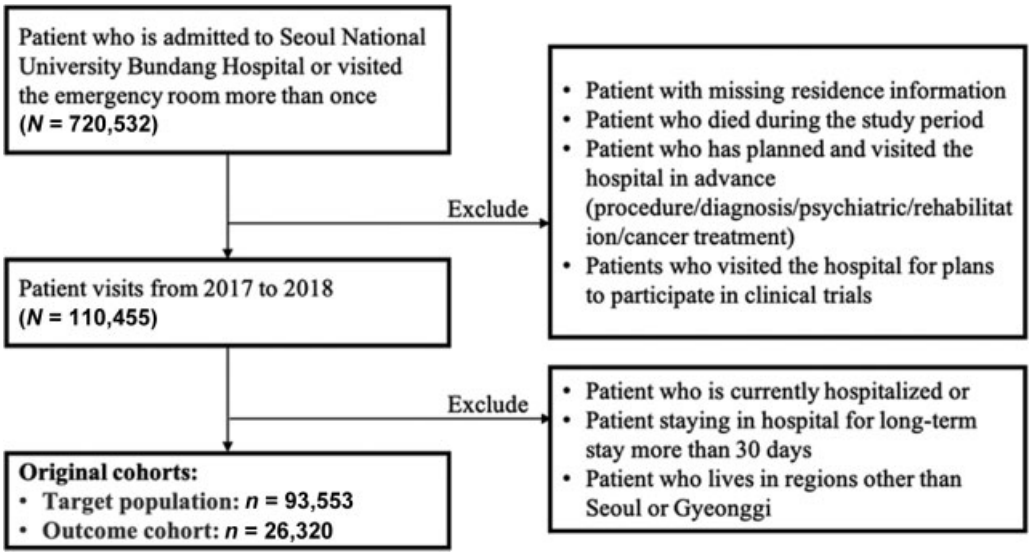

Fig. 1 SNUBH cohort design of the study. SNUBH, Seoul National University Bundang Hospital.

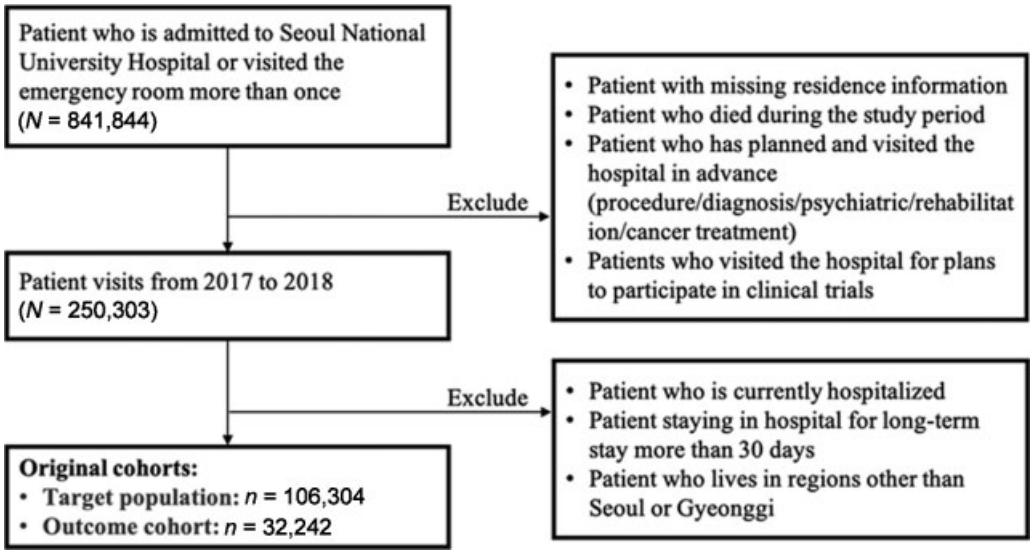

Fig. 2 SNUH cohort design of the study. SNUH, Seoul National University Hospital.

"CONDITION_GROUP_ERA" is composed of a higher hierarchy concept for the given covariate per CONDITON_ERA table. For instance, "Pneumonia" may be a parent of the different subtypes of "pneumonia" that are found in condition era; thus, a person with different diagnosis subtypes of pneumonia will be counted only once in the "pneumonia" group, while without this grouping, this person would contribute a count to each pneumonia subtype the person was diagnosed with. Likewise, "DRUG_ERA" is defined as a span of time when the patient is assumed to be exposed to a particular active ingredient. Notably, "DRUG_ERA" is not the same as a "DRUG_EXPOSURE": exposures are individual records corresponding to the source when the drug was delivered to the patient, while successive periods of "DRUG_EXPOSURE" are combined under certain rules to produce continuous drug eras. These diagnoses and medication records are mainly mapped into SNOMED CT and RxNorm standard terminologies.

- *For surgery and clinical examination tests*: Patient's surgical record was derived from the "PROCEDURE" table in OMOP CDM. In the case of patient clinical examination test data, test order and its result value of a record were extracted from two tables ("MEASUREMENT" and "OBSERVATION" data tables in the database) according to the type of each test.
- *Visit records*: The "VISIT_OCCURRENCE" table includes information about a patient's EHR, either as inpatient, outpatient, or emergency department visits. The number of visits was counted based on the patient's visit type and used as a variable.
- Other demographic information and clinical scores such as patient gender, age group, and location were extracted from the "PERSON" table. Clinical index scores, such as CCI, diabetes complications severity index, and CHA2DS2-VASc score for estimating the risk of stroke of each individual patient, were also used as model variables.

We considered the time boundaries of each feature as follows: (1) long-term covariate combination contains variables acquired during the 365 days prior to discharge date; (2) medium-term setting contains variables included during the 180 days prior to discharge date; and (3) short-term covariate combination contains variables included during the 30 days prior to discharge date.

Unplanned Hospital Readmission

The primary outcome was 30-day unplanned hospital readmission. This was defined as a hospitalization through outpatient visit or emergency room visit in a study period, except planned schedule. We referred to the quality measure of Hospital-Wide All-Cause Unplanned Readmission (HWR) from Centers for Medicare and Medicaid Services (CMS). According to the HWR measure, CMS classifies planned readmissions as planned disease or treatment groups, including chemotherapy, organ transplant, rehabilitation, and other planned treatments or surgeries. We defined scheduled admissions first, and the remaining admissions were assumed to be unscheduled visits. Among the confirmed hospital admissions during the study period, approximately 20% were unplanned. ►**Figure 3** shows the timeline of patient visits during the 30-day hospital revisit period.

Model Development and Clinical Covariate Combinations

A patient-level prediction model was iteratively developed and validated using SNUBH and SNUH CDM data. Two types of experiments were performed in this study. Through the first experiment, our intent was to find the effective variable time for predicting readmission within 30 days. To predict rehospitalization within 30 days, clinical variables that occurred 365 days before discharge date (long-term covariate span), 180 days before discharge (medium-term covariate span), and entering predictive models for variables

Fig. 3 Patient visit timeline based on readmission definition.

before 30 days (short-term covariate span) were explored in the experiment.

In the second experiment, we tested which variables are effective for rehospitalization within 30 days. For diagnoses and drug variables 30 days prior to the discharge date, we designed variables that differ in levels of granularity in each data. Once the prediction model was developed using SNUH data, external evaluation was applied against the SNUBH data. In the opposite case, a model developed with data from SNUBH was evaluated externally using data from SNUH (see ►**Supplementary Fig. S1**, available in the online version only).

Variables such as the patient's demographic information and clinical index scores, diagnosis, medications, visit frequency records, surgeries, and clinical examination test records were included. ►**Table 1** summarizes the differences in the combination settings for each variable. For convenience, we describe each variable combination with a name: DM-SC ("diagnosis, medication, surgeries, clinical exam") and DM ("diagnosis, medication").

DM-SC covariate combination includes all records of a patient's diagnosis, medications, test orders and their results, procedure, or surgical history. Meanwhile, DM combinations 1, 2, and 3 mainly concern patient diagnoses and medication information with different conceptual levels. The lowest granularity reflects individual diagnostic information for each patient visit date (DM 1). Rather, DM 2 allows aggregation of chronic conditions that require frequent ongoing care, instead of treating each diagnostic information as an independent event. With higher granularity of concepts in DM 3, the aggregations of chronic individual diagnostic eras are

Table 1 Differences in combination settings for each variable

Category	DM-SC	DM 1	DM 2	DM 3
Demographics	Gender, age group, index month	Gender, age group, index month, demographics time in cohort	Gender, age group, index month, demographics time in cohort	Gender, age group, index month, demographics time in cohort
Clinical index score	Charlson index, DCSI, Chads2, Chads2Vasc	Charlson index, DCSI, Chads2, Chads2Vasc	Charlson index, DCSI, Chads2Vasc	Charlson index, Chads2Vasc
Diagnosis	Condition occurrence, distinct condition count	Condition occurrence, distinct condition count	Condition era, distinct condition count	Condition group era, distinct condition count
Medication	Drug exposure, drug era, distinct ingredient count	Drug exposure, distinct ingredient count	Drug era, distinct ingredient count	Drug group era, distinct ingredient count
Visit records	Total count, visit types count	Visit types count	Visit types count	Visit types count
Surgeries	Procedure	Distinct procedure count	Distinct procedure count	Distinct procedure count
Clinical examination test	Observation	Distinct observation count	Distinct observation count	Distinct observation count
	Measurement	Distinct measurement count	Distinct measurement count	Distinct measurement count

Abbreviations: DCSI, diabetes complications severity index; DM, diagnosis, medication; DM-SC, diagnosis, medication, surgeries, clinical exam.

"grouped" together under the ancestor hierarchy of the condition concept found in the CONDITON_ERA table in CDM.

In addition, to examine the differences in the characteristics of readmitted patients, the experiment was conducted by dividing the patients into three groups: all age group, patients 65 years of age and older, and children and adolescents under 18 years of age.

We developed machine-learning–based models based on LASSO logistic regression (LR), decision tree (DT), random forest (RF), adaptive boosting (AdaBoost), and gradient boosting machine (GBM) to predict 30-day readmissions and compared their performance. These machine-learning methods were considered owing to their following characteristics. LR is a common and basic algorithm, which is widely used in disease risk prediction and epidemiology.[32] DT is also used in different areas of medical decision making.[33] RF, as an ensemble algorithm of trees, applies a bootstrap algorithm to extract multiple samples from the training set randomly, and trains the samples with the weak classifier.[34] An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.[35] GBM is a distributed and high-performance gradient lifting framework based on a DT algorithm designed for fast computational time, especially with very large datasets.[36]

Patient-level prediction R package developed by the OHDSI was used to train and test the models. Furthermore, 10-fold cross-validation was primarily used for internal validation. To compare the performances of models applied to the prediction of readmission, we used the area under the receiver operating characteristic curve (AUROC) as the primary evaluation criterion. All the models were externally validated to examine their portability.

Table 2 Basic characteristics of SNUH data per visit type

Characteristic		Entire cohort	Derived cohort		*p*-Value
			Readmitted	Not readmitted	
Age, y, mean (SD)		46.8 (27.5)	49.2 (25.8)	45.1 (28.4)	
Gender	Male, *n* (%)	50.4	51.1	51.1	0.001
	Female, *n* (%)	49.6	48.9	48.9	
Age at hospital visit	10 under	18.5	14.5	20.3	
	10s	5.8	5.9	5.7	
	20s	5.6	6	5.4	
	30s	6.7	7.3	6.5	
	40s	7.8	8.5	7.5	
	50s	12.5	14.3	11.8	
	60s	17.5	18.5	17	
	70s	17.1	16.8	17.2	
	80s	7.7	7.2	7.8	
	90s	0.8	0.7	0.9	
Season at time of discharge	Spring	24.9	24.8	26.3	
	Summer	26.3	26.8	26.8	
	Fall	24.3	24.9	24.0	
	Winter	24.5	23.4	25.0	
Admission weekday	Monday	17.6	17.5	17.7	
	Tuesday	15.6	15.3	15.7	
	Wednesday	15.4	15.2	15.5	
	Thursday	15.1	15.5	14.9	
	Friday	11.9	13.4	11.3	
	Saturday	9.1	9.7	8.9	
	Sunday	15.2	13.4	16.0	
Average length of stay, mean (SD)		2.5 (4.4)	2.9 (4.9)	2.4 (4.2)	
Charlson comorbidity index, mean		0.21	0.38	0.18	

Abbreviations: SD, standard deviation; SNUH, Seoul National University Hospital.

Ethical Considerations

The study was performed in compliance with the World Medical Association Declaration of Helsinki and Ethical Principles for Medical Research Involving Human Subjects. This study was approved by the SNUBH Institutional Review Board (X-1908–559–901) and SNUH Institutional Review Board (E-2002–002–1097).

Results

Overall, 106,304 index hospitalizations were included in our SNUH study cohort, of which 32,242 resulted in a 30-day readmission (►Table 2). Individuals had a mean age of 46.8 years, and slightly more than half were males. The average length of stay was 2.5 days in the entire cohort and 2.9 days in the readmitted cohort. In the SNUBH cohort, 93,553 index hospitalizations were included in our study cohort, of which 26,320 resulted in a 30-day readmission (►Table 3). Individuals had a mean age of 46.3 years, and slightly more than half were females. The average length of stay was 2.3 days in the entire cohort and 2.8 days in the readmitted cohort. Individuals in the cohort with a 30-day readmission had markedly different socioeconomic and clinical characteristics compared with those not readmitted.

To predict an outcome occurrence considering time boundaries of features, we developed models with data from two hospitals and compared the performance (►Table 4). In this experiment, no external verification was performed, only the comparison of the results of the two hospital models for the time variable was performed. Here, each model was developed using SNUH data (mother hospital) and SNUBH data (child hospital).

Overall, the performance of the model developed with the data of the parent hospital (SNUH) was high. The model with the best performance in common was GBM, and the LASSO LR and DT showed low performance. Through this experiment, we observed that the effective variable time boundary for readmission within 30 days is to use the clinical variable

Table 3 Basic characteristics of SNUBH data per visit type

Characteristic		Entire cohort	Derived cohort		*p*-Value
			Readmitted	Not readmitted	
Age, y, mean (SD)		46.3 (27.7)	49.2 (25.8)	45.1 (28.4)	
Gender	Male, *n* (%)	49.6	48.8	49.9	0.003
	Female, *n* (%)	50.4	51.2	50.1	
Age at hospital visit	10 under	18.8	13.3	20.9	
	10s	5.4	5.3	5.4	
	20s	4.8	4.7	4.9	
	30s	7.4	7.6	7.3	
	40s	10.4	12.6	9.6	
	50s	13.2	14.5	12.6	
	60s	14.8	15.9	14.3	
	70s	15.6	16.4	15.2	
	80s	8.5	8.5	8.5	
	90s	1.1	1.2	1.1	
Season at time of discharge	Spring	25.7	26.3	25.5	
	Summer	26.4	26.8	26.3	
	Fall	24.2	24.9	23.9	
	Winter	23.7	22.0	24.4	
Admission weekday	Monday	16.7	16.7	16.7	
	Tuesday	14.8	15.0	14.7	
	Wednesday	15.2	15.4	15.1	
	Thursday	14.7	14.7	14.7	
	Friday	14.0	15.0	13.5	
	Saturday	10.3	10.2	10.4	
	Sunday	14.3	13.0	14.8	
Average length of stay, mean (SD)		2.3 (4.3)	2.8 (4.8)	2.1 (4.1)	
Charlson comorbidity index, mean		0.28	0.39	0.26	

Abbreviations: SD, standard deviation; SNUBH, Seoul National University Bundang Hospital.

Table 4 Overall performance on covariate time-boundary settings

Model	Long-term (−365 days)		Medium-term (−180 days)		Short-term (−30 days)	
	Train/test SNUH	Train/test SNUBH	Train/test SNUH	Train/test SNUBH	Train/test SNUH	Train/test SNUBH
LASSO logistic regression	70.85	65.84	70.98	68.64	80.47	76.62
Decision tree	72.03	59.35	72.02	65.65	80.94	74.4
Random forest	**74.02**	65.05	74.13	70.55	**82.75**	78.08
AdaBoost	73.03	64.85	73.16	67.10	81.01	75.64
Gradient boosting machine	71.11	**67.49**	**75.44**	**71.07**	82.52	**79.75**

Abbreviations: SNUBH, Seoul National University Bundang Hospital; SNUH, Seoul National University Hospital.
Note: The highest performance in each column is marked in bold.

for short-term covariate span (30 days before discharge). This result was based on variables such as data from surgery, procedures, and clinical examination result, as well as diagnostic records and medication data. The model performance for the short-term variables was the best. It can thus be interpreted that to predict readmission within 30 days, the model developed with data for 30 days based on the discharge date makes the most accurate prediction. The AUROC curve in ►Fig. 4 shows the results for the best performance described in ►Table 6, SNUH-developed model and short-term covariate combination. It can be seen in ►Fig. 4 that GBM showed the best performance with an AUROC of 84.14, while LASSO LR achieved an AUROC of 80.14.

Hence, in our second experiment, we developed models for variables prior to 30 days before discharge and performed external validation of each hospital model by using data such as diagnoses and drugs, which are relatively unlikely to have gaps in term mapping (DM covariates setting). ►Table 5 shows the results of the SNUBH data-trained model, with external validation (AUROC, with short-term covariates), and ►Table 6 shows the results of the SNUH data-trained model. In DM 3, the variables were calculated with Group Era (higher concept of period variable) to include the higher concept, while DM 1 contains only individual diagnoses and medications.

We categorized the entire experimental cohort data into three groups: all age group, patients 65 years of age and

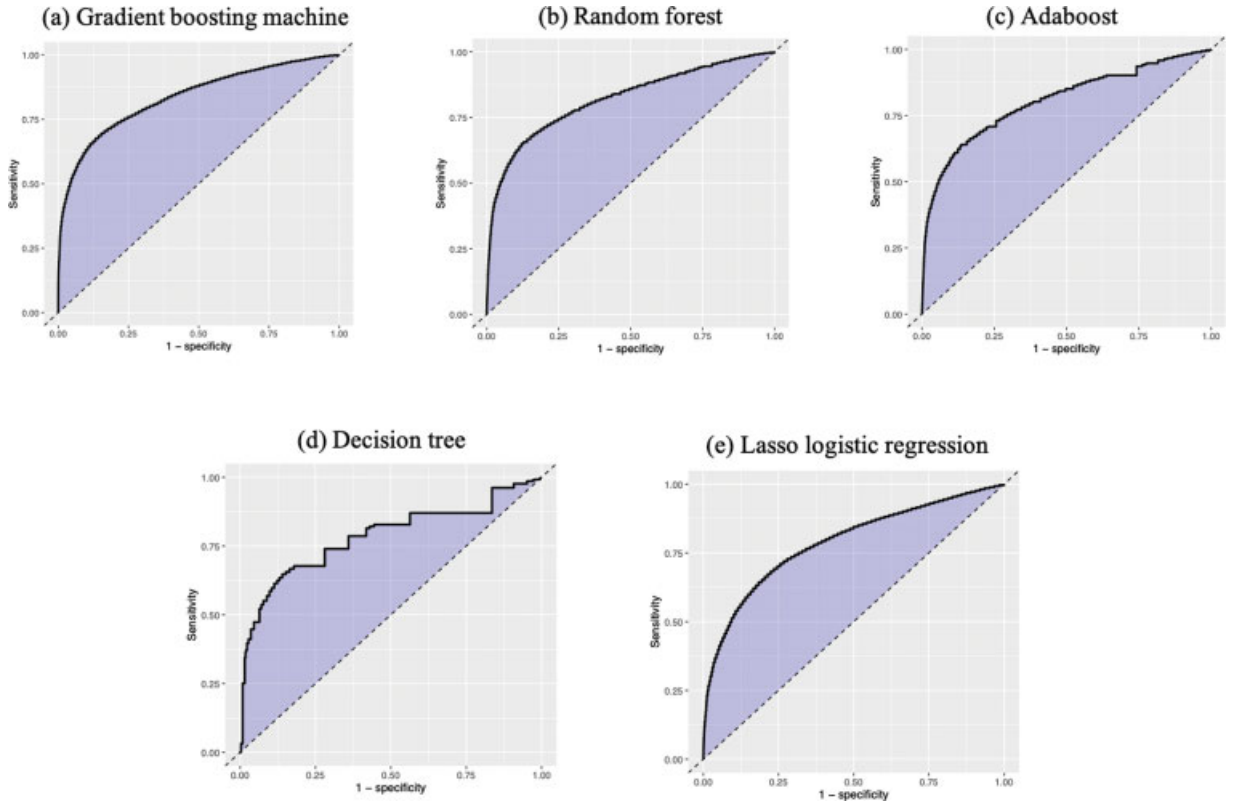

Fig. 4 Results of the developed model based on SNUH test data with short-term covariates. SNUH, Seoul National University Hospital.

Table 5 Overall performance on SNUBH-data-trained classification models

Model	DM 1		DM 2		DM 3	
	Train/test SNUBH	Validation SNUH	Train/test SNUBH	Validation SNUH	Train/test SNUBH	Validation SNUH
LASSO logistic regression	65.84	62.44	68.64	65.22	76.62	73.38
Decision tree	64.03	59.35	65.65	62.44	74.40	70.37
Random forest	65.05	**65.66**	70.55	**67.87**	78.08	75.10
AdaBoost	64.85	61.13	67.10	64.39	75.64	73.93
Gradient boosting machine	**67.49**	63.75	**71.07**	65.73	**79.75**	**75.34**

Abbreviations: DM, diagnosis, medication; SNUBH, Seoul National University Bundang Hospital; SNUH, Seoul National University Hospital.
Note: The highest performance in each column is marked in bold.

Table 6 Overall performance on SNUH-data-trained classification models

Model	DM 1		DM 2		DM 3	
	Train/test SNUH	Validation SNUBH	Train/test SNUH	Validation SNUBH	Train/test SNUH	Validation SNUBH
LASSO logistic regression	77.04	72.12	79.07	73.70	80.17	76.17
Decision tree	75.87	71.66	80.76	74.56	81.01	73.05
Random forest	79.33	74.97	82.36	**77.34**	82.24	77.66
AdaBoost	77.11	73.36	81.08	76.49	81.29	78.14
Gradient boosting machine	**80.90**	**75.94**	**83.90**	76.71	**84.14**	**78.33**

Abbreviations: DM, diagnosis, medication; SNUBH, Seoul National University Bundang Hospital; SNUH, Seoul National University Hospital.
Note: The highest performance in each column is marked in bold.

older, and children and adolescents under 18 years of age, based on the age at the time of the patient visit. When comparing the performance of the prediction model based on the age of the patient, it was confirmed that the prediction performance was lowest in the group of children and adolescents under 18 years of age (►Fig. 5). From this result, it can be assumed that the predictive model we developed was made suitable for predicting elderly patients.

Discussion

In this study, we demonstrated how prediction tools can be integrated generically into two different clinical settings and provide an exemplary use case for predicting 30-day hospital readmission. We compared the performance of models developed using data from two hospitals and compared the prediction performance of readmissions.

Conventional patient readmission risk assessments are performed using a variety of assessment tools ranging from multidisciplinary patient interviews to simple screening tools with fewer variables.[8,10,37–39] Many of the previously developed readmission prediction models, including the readmission prediction scores mentioned above, account for most of the models based on statistical calculations or LR analysis using several clinical variables. With the development of machine-learning technology, attempts to introduce new machine-learning techniques in predicting readmission are increasing.[12,13,40,41] However, there are insufficient proven cases applicable to actual clinical environments. From this viewpoint, the development of a CDM-based prediction tool has an

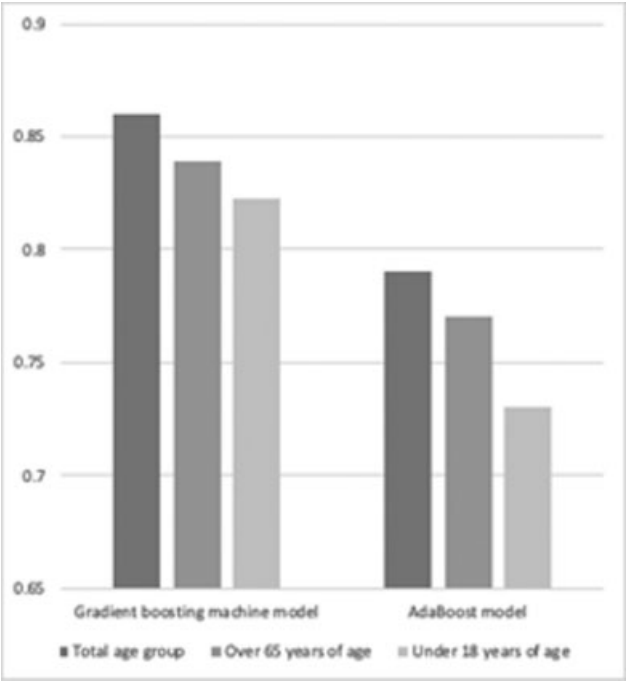

Fig. 5 The result of the GBM model and the AdaBoost model by patient age group. AdaBoost, adaptive boosting; GBM, gradient boosting machine.

advantage. The prediction model using CDM is easily transferable and can be applied to various institutions that have CDM data. Because each data element constituting the CDM data is mapped to standard terminologies, it is possible to interpret the common meaning of each institutional model.

By comparing the results of the prediction models applied to the two hospitals based on OMOP CDM in this study, we corroborate that the type of patient visit had the greatest influence on the prediction of hospital readmission within 30 days. In the performance comparison experiment of the model made from the data of the parent hospital and the child hospital, there may be various causes such as differences in patient composition of each hospital or differences in term mapping in the process of converting source data to CDM data. In our first experiment, the results were based on variables such as data from surgery, procedures, and clinical examination result, as well as diagnostic records and medication data (DM-SC covariates). The reason for the difference in the performance of the two hospital models is probably due to the distribution of the source data or the mapping of terms that are different for each domain of the CDM. Among them, the prediction performance was the best in the short-term covariates input as patient variables within 30 days from the discharge date. It can thus be interpreted that to predict readmission within 30 days, the model developed with data for 30 days based on the discharge date makes the most accurate prediction.

We confirmed how to combine the semantic units of variable data used when constructing the readmission prediction model that shows good results. Data from only two hospitals were used for this study, and there is a limitation in not being able to use data from several hospitals that have OMOP CDM. However, it can be applied to other hospitals as well. Furthermore, with more sophisticated data processing, such as adopting deep learning-based techniques, the model can be expected to perform better. We hope to derive and apply more features of clinical entities or deep learning techniques in future studies.

Conclusions

In this study, predictive models were developed based on CDM that could explore clinical variables to predict hospital readmission within 30 days. As a result, in the model targeting the 30-day prediction, when the data 30 days before the discharge date were used, the prediction performance was the best. In addition, it was confirmed that making a predictive model by creating a variable with data on a high-level concept yields a better performance. The CDM prediction models showed external validity in terms of the model performance as well as transportability. The model developed in this study can be expanded and be used by clinicians in the field.

Note

The CDM-based prediction model has the following advantages. It can be easily reintegrated when migrating to a different EHR with analysis code adoption, either as an embedded frame in the EHR or as a standalone application. In addition, it can be easily expanded to another hospital based on OMOP CDM, which could be easily transferred and further developed with regard to our approach.

Authors' Contributions

B.R. analyzed the data and drafted the manuscript as the first author. S.K. helped prepare and evaluate the data. S.Y. helped analyze the data and managed the overall study, and J.C. supervised the overall study.

Funding

This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program) (20004927, Advancing and expanding CDM-based distributed biohealth data platform) funded by Korea's Ministry of Trade, Industry, and Energy.

Conflict of Interest

None declared.

References

1 Toomey SL, Peltz A, Loren S, et al. Potentially preventable 30-day hospital readmissions at a children's hospital. Pediatrics 2016; 138(02):e20154182
2 McIlvennan CK, Eapen ZJ, Allen LA. Hospital readmissions reduction program. Circulation 2015;131(20):1796–1803
3 Hong J, Choi K, Lee J, Lee E. A study on the factors related to the readmission and ambulatory visit in an university hospital: using patient care information DB. J Korean Soc Med Informatics 2000;6 (04):23–33
4 Zuckerman RB, Sheingold SH, Orav EJ, Ruhter J, Epstein AM. Readmissions, observation, and the hospital readmissions reduction program. N Engl J Med 2016;374(16):1543–1551
5 Thompson MP, Waters TM, Kaplan CM, Cao Y, Bazzoli GJ. Most hospitals received annual penalties for excess readmissions, but some fared better than others. Health Aff (Millwood) 2017;36(05):893–901
6 Hasan O, Meltzer DO, Shaykevich SA, et al. Hospital readmission in general medicine patients: a prediction model. J Gen Intern Med 2010;25(03):211–219
7 Nguyen OK, Makam AN, Clark C, et al. Predicting all-cause readmissions using electronic health record data from the entire hospitalization: model development and comparison. J Hosp Med 2016;11(07):473–480
8 Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: a systematic review. JAMA 2011;306(15):1688–1698
9 Choudhry SA, Li J, Davis D, Erdmann C, Sikka R, Sutariya B. A public-private partnership develops and externally validates a 30-day hospital readmission risk prediction model. Online J Public Health Inform 2013;5(02):219
10 Silverstein MD, Qin H, Mercer SQ, Fong J, Haydar Z. Risk factors for 30-day hospital readmission in patients ≥65 years of age. Proc (Bayl Univ Med Cent) 2008;21(04):363–372
11 Swain MJ, Kharrazi H. Feasibility of 30-day hospital readmission prediction modeling based on health information exchange data. Int J Med Inform 2015;84(12):1048–1056
12 Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. npj. Digit Med 2018;1(01):1–10
13 Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. JAMA Cardiol 2017;2(02):204–209

14 Maali Y, Perez-Concha O, Coiera E, Roffe D, Day RO, Gallego B. Predicting 7-day, 30-day and 60-day all-cause unplanned readmission: a case study of a Sydney hospital. BMC Med Inform Decis Mak 2018;18(01):1–11
15 Robinson R, Hudali T. The HOSPITAL score and LACE index as predictors of 30 day readmission in a retrospective study at a university-affiliated community hospital. PeerJ 2017;5(03):e3137
16 Miller WD, Nguyen K, Vangala S, Dowling E. Clinicians can independently predict 30-day hospital readmissions as well as the LACE index. BMC Health Serv Res 2018;18(01):32
17 Shameer K, Johnson KW, Yahi A, et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: A case-study using Mount Sinai Heart Failure Cohort. Pac Symp Biocomput 2017;22:276–287
18 Blacketer C. Chapter 4: The Common Data Model. In: The Book of OHDSI. Observational Health Data Sciences and Informatics. Accessed May 31, 2021 at: https://ohdsi.github.io/TheBook OfOhdsi/CommonDataModel.html
19 Williams R, Kontopantelis E, Buchan I, Peek N. Clinical code set engineering for reusing EHR data for research: a review. J Biomed Inform 2017;70:1–13
20 Sentinel Initiative. Homepage. Accessed May 31, 2021 at: http://www.mini-sentinel.org/
21 The National Patient-Centered Clinical Research Network. Data. Accessed May 31, 2021 at: https://pcornet.org/data/
22 Ross TR, Ng D, Brown JS, et al. The HMO research network virtual data warehouse: a public data model to support collaboration. EGMS (Wash DC) 2014;2(01):1049
23 OMOP CDM v6.0. Accessed May 21, 2021 at: https://ohdsi.github.io/CommonDataModel/cdm60.html#OMOP_CDM_v60
24 CDISC. SDTM. Accessed May 31, 2021 at: https://www.cdisc.org/standards/foundational/sdtm
25 Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016;64:333–341
26 Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(01):54–60
27 FitzHenry F, Resnic FS, Robbins SL, et al. Creating a common data model for comparative effectiveness with the Observational Medical Outcomes Partnership. Appl Clin Inform 2015;6(03):536–547
28 OHDSI – Observational Health Data Sciences and Informatics. Accessed December 29, 2020 at: https://www.ohdsi.org/
29 Park RW. Sharing clinical big data while protecting confidentiality and security: observational health data sciences and informatics. Healthc Inform Res 2017;23(01):1–3
30 Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574–578
31 OMOP CDM Specifications. Accessed June 29, 2021 at: https://ohdsi.github.io/CommonDataModel/
32 Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002;35(5–6):352–359
33 Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision trees: an overview and their use in medicine. J Med Syst 2002;26(05):445–463
34 Breiman L. Random forests. Mach Learn 2001;45(01):5–32
35 Freund Y, Schapire RE. A decision-theoretic generalization of online learning and an application to boosting. J Comput Syst Sci 1997;55(01):119–139
36 Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 2017;30:3146–3154
37 Zhou H, Della PR, Roberts P, Goh L, Dhaliwal SS. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. BMJ Open 2016;6(06):e011060
38 Smith DM, Giobbie-Hurder A, Weinberger M, et al; Department of Veterans Affairs Cooperative Study Group on Primary Care and Readmissions. Predicting non-elective hospital readmissions: a multi-site study. J Clin Epidemiol 2000;53(11):1113–1118
39 Robinson R. The HOSPITAL score as a predictor of 30 day readmission in a retrospective study at a university affiliated community hospital. PeerJ 2016;4(09):e2441
40 Jiang S, Chin KS, Qu G, Tsui KL. An integrated machine learning framework for hospital readmission prediction. Knowl Base Syst 2018;146:73–90
41 Jamei M, Nisnevich A, Wetchler E, Sudat S, Liu E. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. PLoS One 2017;12(07):e0181173