공학석사 학위논문

A Retrospective Study on Cardiovascular Disease Risk Prediction based on Machine Learning Using Electronic Medical Records

전자의무기록을 활용한 머신러닝 기반 심혈관질환 관련 위험 예측 후향적 연구

울산대학교 대학원

의과학과

안임진

A Retrospective Study on Cardiovascular Disease Risk Prediction based on Machine Learning Using Electronic Medical Records

지도교수 김영학

이 논문을 공학석사학위 논문으로 제출함

2022년 08월

울산대학교 대학원

의과학과

안임진

안임진의 공학석사학위 논문을 인준함

심사위원 이계화 (인)

심사위원 김영학 (인) 김영학인

심사위원 전태준 (인)

울산대학교 대학원

2022 년 8 월

Abstract

Background

The clinical data stored in medical institutions is rapidly increasing with the development of healthcare-related technologies. To apply artificial intelligence (AI) approaches to medical data such as electronic medical records (EMRs), it is necessary to establish and curate a specialized database. Specially, cardiovascular diseases (CVDs) are difficult to diagnose early and have risk factors that are easy to overlook. Early prediction and personalization of treatment through the use of AI may help clinicians and patients manage CVDs more effectively. Moreover, since effective resource management in hospitals can improve the quality of medical service, predicting a patient’s hospitalization period may support making judicious decision regarding bed management.

Objectives

First, we aim to build a suitable database (CardioNet) for CVDs that can utilize AI technology, contributing to the overall care of patients with CVDs. Second, we aim to develop a Machine Learning (ML)-based model for predicting the discharge probability and to explain the individual risk factors for improving the patient management. Third, we aim to develop a Deep Learning (DL)-based model for estimated glomerular filtration rate (eGFR) prediction of inpatients with heart failure (HF) and to visualize results of prediction to enhance the quality of medical services.

Methods

First, we build the CardioNet with data from 748,474 patients, which consisted of anonymized records who had visited the Asan Medical Center (AMC) or Ulsan University Hospital (UUH) because of CVDs between January 1, 2000, and December 31, 2016. In addition, we pre-processed EMRs to remove errors and duplications, and performed natural language processing to structuralize the free-text readings. Second, we processed the data to create a suitable dataset by reindexing the date-index, integrating the present features with past features from the previous 3 years, and imputing missing values. Subsequently, we trained the ML-based predictive models, and predicted the discharge probability within 3 days and explained the outcomes of the model by identifying, quantifying, and visualizing its features. Third, we extracted data of hospitalized patients with HF, performed pre-processing and created a dataset including time series to train a DL-based model to make predictions for eGFR. Additionally, we proposed visualized outcomes of the DL-based model for utilizing the results in clinical practices.

Results

CardioNet is a comprehensive database that can serve as a training set for AI models and assist in all aspects of clinical management of CVDs. It comprises information extracted from EHRs and results of readings of CVD-related digital tests. It consists of 27 tables, a code-master table, and a descriptive table.

In order to predict hospital discharge prediction, we experimented with 5 ML-based models using 5 cross-validations. The extreme gradient boosting (XGB), which was selected as the final model, accomplished an average area under the receiver operating characteristic curve score that was 0.865 higher than for other models. Furthermore, we performed feature reduction, represented the feature importance, and assessed prediction outcomes. One of the outcomes, the individual explainer, provides a discharge score during hospitalization and a daily feature influence score to the medical team and patients. Finally, we visualized simulated bed management to use the outcomes.

In order to predict the eGFR to prevent the patients with HF to risk, we performed pre-processing to create sequential learning dataset and developed the DL-based model based on recurrent neural networks. The predictive model we developed learns 24 hours of data, predicts eGFR levels after 12, 24, 36, and 48 hours and predicts one of five risk labels. Our DL-based model achieved the mean squared error of 169.626, the mean absolute error of 5.82, and the accuracy of classification was 85.1%. Subsequently, we visualized the outcomes of models including overall eGFR graphs and divided graphs for each time step.

Conclusions

First, we established the comprehensive database specialized in CVDs. We are actively supporting multi-center research, which may require further data processing, depending on the subject of the study. CardioNet will serve as the fundamental database for future CVD-related research projects. Second, we proposed an individual explainer based on an ML-based predictive model, which provides the discharge probability and relative contributions of individual features. Our model can assist medical teams and patients in identifying individual and common risk factors in CVDs and support hospital administrators in improving the management of hospital beds and other resources. Third, we conducted the effective pre-processing method for generating sequential data from EMRs. We developed the DL-based predictive model providing the value and risk of eGFR for inpatients with HF. Additionally, we presented overall and divided graph by time step which could support the medical team and patients in managing the risk of HF and CVDs in advance.

Keywords: electronic medical records, cardiovascular diseases, artificial intelligence, database, hospital discharge prediction, risk prediction.

Contents

Abbreviation

AI: Artificial Intelligence
AMC: Asan Medical Center
AUROC: area under the receiver operating characteristic
CKD: chronic kidney diseases
CKD-EPI: chronic kidney disease epidemiology collaboration
CT: Computed tomography
CVD: cardiovascular disease
DL: deep learning
eGFR: estimated glomerular filtration
EHR: electronic health record
EMR: electronic medical record
GBM: gradient boosting algorithm
HF: heart failure
ICU: intensive care unit
INDT: date of visitation or admission
INNO: patient encounter number
LOS: length of stay
MDRD: modification of diet in renal disease
ML: machine learning
MRI: Magnetic resonance imaging
OHE: one-hot encoding
OUDT: date of discharge
PAID: patient identification
RFECV: recursive feature elimination with cross-validation
ROC: receiver operating characteristic
SPECT: Single-photon emission computed tomography
XAI: explainable artificial intelligence
XGB: extreme gradient boosting

List of Tables

List of Figures

Introduction

The clinical data stored in medical institutions is rapidly increasing with the development of healthcare-related technologies in recent. The electronic medical records (EMRs) are one of the medical big-data and includes various medical records of the patient [1]. Although patient records in EMRs were not utilized due to privacy protection, a retrospective study could be conducted through de-identification processes such as pseudonymization or anonymization. A retrospective study using EMRs can perform various risk prediction studies and can be used as Real-World Evidence [2].

The cardiovascular disease (CVD) is one of the acute and chronic diseases accompanied by several comorbidities, and requires continuous and active management. For this purpose, there are varied artificial intelligence (AI)-based studies including machine learning (ML) and deep learning (DL) algorithms [3-5]. The following three studies were designed and conducted to support the management of patients with CVDs. First, we aimed to construct the database specialized in CVDs (CardioNet) that can be continuously used for future CVDs research. Second, we aimed to develop a ML-based model to perform CVD-related predictive studies using the constructed database. The purpose of this study is to support efficient utilization of hospital resources by predicting the discharge of inpatients. Third, we aimed to develop a DL-based model using the identical database and predict the estimated glomerular filtration rate to detect the risk of heart failure (HF) of hospitalized patients.

First, clinical data related to CVDs that require active management include basic outpatient and hospitalization data, as well as various unusual tests such as echocardiography and exercise stress test. In order to help carry out CVD-related medical informatics research by integrating these structured and unstructured data, we built the database specializing in CVDs. We extracted anonymized data, removed the error data and outliers according to clinically plausible standards. In addition, we structuralized free-text readings through natural language processing. The established database could increase the usability of EMRs and support secondary derivative research.

Second, we developed a ML-based model by extracting CVD-related inpatient data from the established CardioNet. This study predicted the discharge probability of inpatients with CVDs, visualized individual risk factors and probabilities of discharge through personalized explainer. Consequently, we suggested that this study could help improve hospital process management by visualized simulation.

Third, we obtained the data of inpatient who had hospitalized for HF from the CardioNet, and converted EMRs into time series datasets that DL-based model can learn. Subsequently, we proposed the DL-based model that can predict the estimated glomerular filtration rate (eGFR) and detect the risk in patients with HF and/or chronic kidney disease (CKD). In addition, we presented the prediction results of eGFR after 12, 24, 36, and 48 hours as a plots, suggesting that this study could be utilized in clinical practice.

In conclusion, we built a database that could be efficiently used for retrospective research, developed ML and DL-based predictive models, and proposed visualized outcomes. The results of the models could assist medical teams and patients with CVDs, including HF, in continuously and actively managing them, and are expected to be used in clinical practice by providing prediction results. Our ML and DL-based models in this study could contribute to precision medicine and digital healthcare through the expansion of risk prediction research in the future.

The first study was published as "CardioNet: a manually curated database for artificial intelligence-based research on cardiovascular diseases" in BMC Medical Informatics and Decision Making on January 28, 2021. The second study was also published as "Machine Learning–Based Hospital Discharge Prediction for Patients With Cardiovascular Diseases: Development and Usability Study" in JMIR medical informatics on November 17, 2021.

Chapter 1. CardioNet: A manually curated database for artificial intelligence-based research on cardiovascular diseases

Introduction

background

Cardiovascular diseases (CVDs) are disorders related to the heart and blood vessels responsible for the blood supply in the human body. According to the World Health Organization (WHO), every year, an estimated 17 million people globally die of CVDs, particularly heart attacks and strokes [3]. CVDs are often caused by environmental factors, such as obesity, smoking, drinking, and stress, or genetic factors, such as underlying disease and family history. CVDs require relatively intense management, and the morbidity of long-term complications is high. In particular, metabolic diseases such as hypertension, diabetes, and dyslipidemia are typical complications. Most of these diseases have no serious symptoms at their onset and are difficult to diagnose, making it easy to overlook the risk or severity of CVDs. Therefore, it is necessary to analyze the CVD-related clinical data to identify risk factors and to develop a predictive model that can help clinicians accurately diagnose the diseases early through consideration of individual characteristics of patients.

Recent advances in artificial intelligence (AI) technology have enabled early detection of several diseases and have already shown performance that approximates or exceeds that of a physician [6, 7]. Deep learning technology, such as the Convolutional Neural Network and Recurrent Neural Network, has shown great results in the analysis of raw medical image and signal data [8, 9]. Conversely, the performance of AI technology using structured medical data stored in electronic health records (EHRs) does not reach the effectiveness seen in image and signal analyses [10, 11]. AI approaches likely excel in medical imaging and signal examinations because of the specialized application of these modalities in the diagnoses of distinct diseases and their use for an explicit purpose. When a disease is diagnosed or suspected, there are inherent representative signs or patterns. The advanced abilities of deep learning to analyze images and signal data rely on the ability to segment given data, identify and learn features directly from the data with minimal external manipulation, and accurately distinguish the essential and inherent features.

For the patient-specific evaluation of CVDs, a database optimized for CVDs based on EHR should be established. EHRs have no specific purpose; they collect and store basic medical information cumulatively such as records of encounters, physical measurements, diagnoses, and medications. Furthermore, unstructured data containing medical information that is relatively important for diagnosis of CVDs include image and signal examinations such as echocardiography, coronary artery computed tomography (CT), electrocardiography (ECG), and cardiac stress test. Since unstructured readings are written in text, it is difficult to use them to train the AI models. Once these unstructured data have undergone proper pre-processing, they should also be integrated into the database to allow for intensive and multi-faceted CVD-related AI research.

Constructing an advanced systematic database allows AI technology to be used efficiently to help clinicians and patients make decisions at each stage of clinical care. In particular, since CVDs require long-term and active management, it is essential for AI to utilize the patients' characteristics, rather than existing diagnostic-based systems. The schema for the specialized cardiovascular database (CardioNet) is shown in Figure 1.

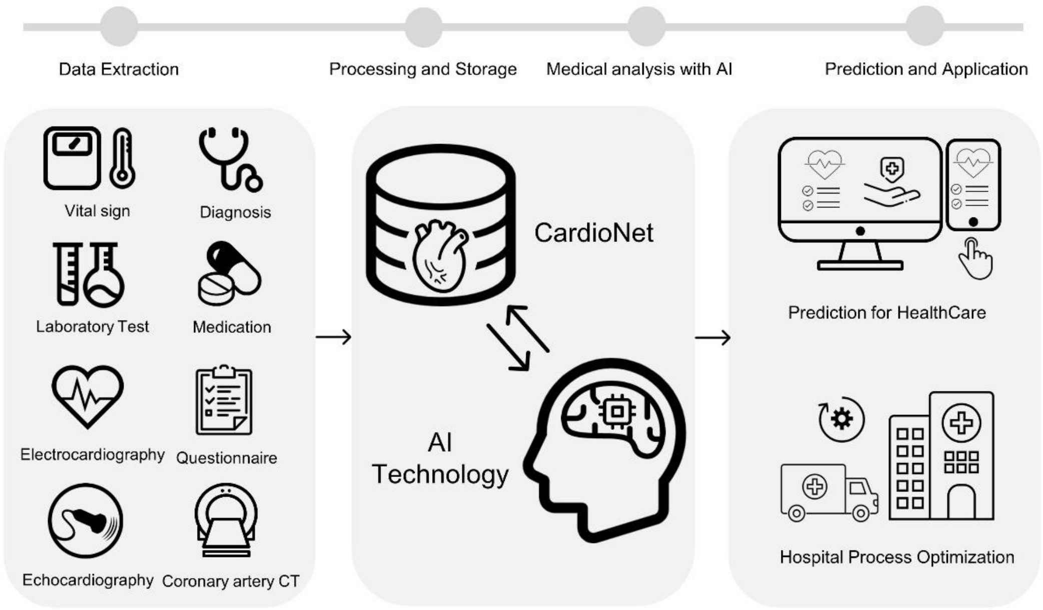


Figure 1 Overview of CardioNet

The purpose of this study was to build CardioNet for CVDs that will allow AI technology to be applied to clinical “big data” based on EHRs. Through this work, we expect to contribute to the discovery of risk factors and detection of their interactions, with the ultimate goal of preventing disease progression and improving treatment planning by early prediction of the occurrence of CVDs and better management of prognosis in overall care for CVDs.

The data used in this study are the records of patients who visited the Asan Medical Center (AMC) in Seoul, Korea, or Ulsan University Hospital (UUH) in Ulsan, Korea, between January 1, 2000, and December 31, 2016. Data were collected from patients diagnosed with heart disease or suspected of having heart disease at the emergency room (ER) or the Departments of Cardiology or Thoracic Surgery or have undergone CVD-related examinations such as echocardiography, ECG, coronary artery CT, and treadmill stress tests.

The main contribution of this work can be summarized in the steps used to establish and optimize CardioNet. First, we integrated unstructured data such as readings of medical examinations with structured data based on EHRs in this study. We built a large-scale and comprehensive clinical database that is suitable for use within diverse medical AI research. Currently, most clinical AI research is conducted as independent studies in specific areas of imaging and signals. However, patients’ clinical outcomes and data should be analyzed with an integrated approach. CardioNet is a database that contains key information about CVDs and can be used as a training data set for AI models on a number of CVD-related topics.

Second, we performed natural language processing (NLP) to structuralize the unstructured information, refining the data to allow for immediate application of AI. Specifically, since the results of CVD-related imaging or signal examinations are mostly written in free-text, NLP was needed to improve the performance of AI, which depends on the pre-processing of the data. In this process, the clinical knowledge naturally included in the data can increase reliability.

Third, we standardized the data in the CardioNet for convergent multi-center research to ensure interoperability. The structure and rules of the EHRs and the coding system are different at each hospital, making multi-institutional research difficult. Standardized code systems have been proposed and utilized for a long time, and there is a trend to take advantage of the globally integrated code system standards that unify data structures, such as the common data model (CDM). However, because the standardization system for the results of various tests, such as imaging for CVDs, is not yet included in the CDM, we created our own variables for them. It will be possible to convert the system quickly when the relevant standardization is established.

Methods

The overall process for building CardioNet is depicted in Figure 2. There are a total of five steps involved: data extraction, structurization, cleansing, standardization, and validation. A detailed description of each step is as follows.

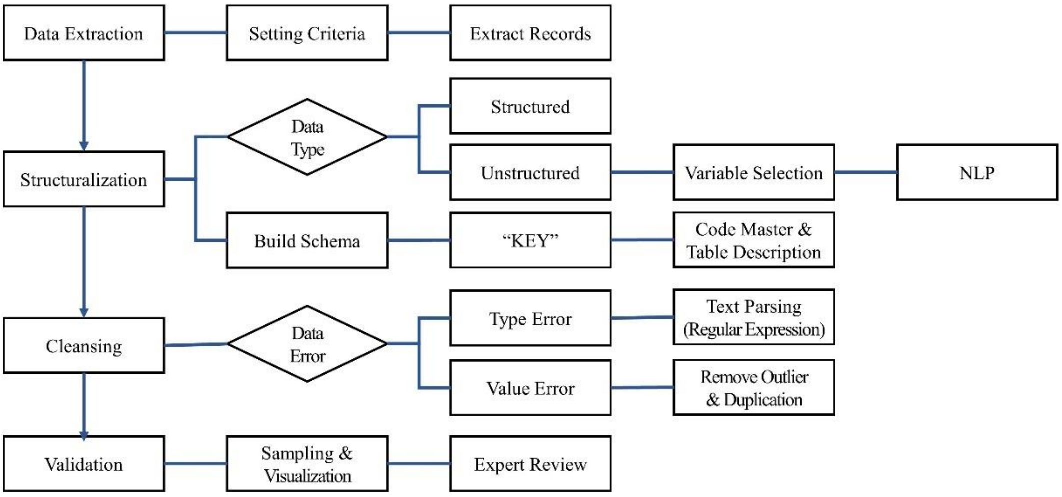


Figure 2 The overall process of building the CardioNet

Description of patients

Anonymity of data

The collection of data and data preparation received AMC and UUH institutional review board approval (IRB) with waived informed consent. It is mandatory for all researchers to protect patients' privacy and to make sure that the data cannot be traced back. AMC has a system called "ABLE (Asan BiomedicaL Environment)", in which only authorized researchers to review anonymous sample data and to extract data after IRB approval. Also, the de-identification and extraction process in this system are conducted through the IT service management unit of AMC's medical information office and the honest broker of the research information unit, not the research participants. UUH's data was similarly anonymized. The list of de-identified information in line with the health insurance portability and accountability act (HIPAA) and hospital policy is shown in Table 1.

Table 1 The list of De-identified data
*Including medical personnel **"Date of birth" is not personally identifiable information; IRB approval is required if information up to "date" is required

No.	De-identified information
1	Unique identification information (resident/alien registration number, passport number)
2	Names (including Chinese characters, English name, pen name, etc.)*
3	Detailed address (detailed address below eup/myeon/dong)
4	All phone numbers (including mobile phone/home/company/fax number)
5	E-mail addresses
6	Medical record number
7	Patient registration number
8	Health insurance card number, Welfare recipient number
9	Accounts number, Credit card number
10	Certificate/License number, Student number
11	Vehicle number, registration number & serial number of various devices
12	Full-face photographs or equivalent (still photo, video, CCTV, video)
13	Identification code (member ID, employee number)
14	IP (Internet Protocol) address, Mac (Media Access Control) address
15	URLs (Universal Resource Locators)
16	Biometric identifiers: fingerprint, iris, vein, voice, handwriting, personally identifiable genetic information
17	Any other personally identifiable information (pathological number)
18	Date of birth**
19	Any other unique identifying information (military number, registration number of the individual business operator)
20	The indirect identification information contained in the information collection is also deleted in principle if it is not related to the purpose of data use.

Detail of data

Data related to all patients who had visited AMC or UUH for CVDs or related complications were collected for this study. We obtained the anonymized records of 748,474 patients who had visited AMC or UUH, from January 1, 2000, to December 31, 2016. Data were obtained from 572,811 patients seen at AMC and 175,663 patients seen at UUH within the same time period. Because the depth of information stored and retained by each hospital is different, the variables were based on the relatively detailed AMC records. In order to prepare to build the CardioNet, we set the following specific criteria for inclusion of patient data:

- Patients who had visited the Departments of Cardiology or Thoracic Surgery.
- Patients who had visited the ER and were assigned International Classification of Diseases, 10th version (ICD-10) codes related to CVDs.

- The codes I00-I99 were related to the diseases of the circulatory system, while R00-R03, R06, R068, R073, and R074 were related to symptoms and signs involving the circulatory and respiratory systems.
- Patients who had undergone coronary artery CT as part of their health screening procedures.
- Patients who had undergone one of the following clinical examinations: thallium single-photon emission computed tomography (SPECT), 2D-echocardiography, treadmill test, and Holter monitoring test.

Data extracted

We obtained data from the EHRs, order communication system, and picture archiving and communication system (PACS). These medical record systems contain information such as patient demographics, vital signs, encounters (e.g., inpatient/outpatient visits, ER visits, and health screening), physical measurements, diagnosis, surgery (order, schedule, and summary), digital medical tests (CT, magnetic resonance imaging [MRI], X-ray, coronary angiography, echocardiography, Holter monitoring, treadmill tests, etc.), laboratory tests (pathology order and result), medication (order, prescription, and history), procedures (order and materials), blood transfusion (order and result), human-derived materials, patient history questionnaire (personal and family medical history, lifestyle and habits), and billing and claim history. By using the service of a pre-built clinical data warehouse, we extracted structured and unstructured data separately, with all data undergoing the process of de-identification.

Data extracted for the establishment of CardioNet included the following:

- Demographics: date of birth, sex, national code, address, blood type (ABO, RH), death date, and death date of a cancer patient (one row per patient).
- Vital signs: measurement time and date, reason for absence of measurement, body temperature, blood pressure, respiratory rate, pulse, oxygen saturation, and consciousness status (one row for each patient seen in the ER).
- Physical information: age, height, weight, blood pressure, pulse, respiratory rate, body mass index (BMI), body surface area, and measurement date (one row per encounter).
- Visits: date of visit, date of admission and discharge, type of visit, medical department, hospitalization, duration of stay in the intensive care unit, type of discharge, and the result of treatment (one row per encounter).
- Diagnosis: date of diagnosis, type of visit, medical department, and ICD-10th code (one row per encounter).
- Surgery: date of admission and discharge, date of surgery or treatment, sequential number of surgery, surgery type (before/after the surgery), diagnosis (before/after the surgery), surgery category, surgical department, and method and time of anesthesia (one row per encounter).
- Digital tests: date of visit, age, department of examination, code of examination, date of order, and reports or readings of the result. (one row per test).

- Laboratory test result: code of pathology examination, number of work, test result, and unit of result (one row per test).
- Medication: medical department, type of visit, date of prescription, code of prescription, active ingredient in medication, indication, category of medicinal effect, and duration of treatment (one row per encounter).
- Procedure: medical department, date of order, code of order, time of order, material code, capacity of materials, and place of patient and materials (one row per procedure).
- Blood transfusion: date of order, code of order, ordering department, sequential number of blood, quantity of prescribed/released blood, and time released (one row per order).
- Human-derived material: date of extraction, code of diagnosis, name of diagnosis, tissue sample description (status and amount of cancer/normal, plasma/buffy coat, type of organ), and information on bone marrow (status and amount of cerebrospinal fluid/bone marrow/blood stored) (one row per patient).
- Patient history: marital status, religion, education, exercise habits over the last three months, lifestyle and habits information (e.g., alcohol, smoking), and personal and family medical history (one row per encounter).

Data processing

Before pre-processing, a Primary Key was generated to connect the data. All data, except demographics and information on human-derived materials, include the patient ID (PAID) and the patient's encounter number (INNO) columns. KEY column was created by concatenating the PAID and INNO to each table to connect all data. Demographics and information on human-derived materials are unique contents that are modified only when the information changes. For this reason, they do not have INNOs, but are linked to other data with PAID. As an example, a patient with the PAID of 100 who visits the hospital for the first time would have the KEY of '100_1'.). In the existing in-hospital medical record system, it was inconvenient to extract all data satisfying specific conditions. We were able to extract data easily and quickly by connecting all data through the KEY.

Structured data

Most of the extracted structured data were quantitative in nature with simple and formal structures, making the pre-processing relatively simple. However, some data required further processing through the removal of data errors and outliers based on clinical knowledge, such as clarifying the meaning of each result and comparing the value to the normal range. Selected aspects of this processing are described below.

Physical information and vital signs Physical information including body measurements, vital signs such as height, weight, blood pressure, pulse, and respiratory rate, are continuous variables, so we identified the distribution of each and removed the implausible data judged to be errors and outliers from a clinical point of view. The following criteria were used:

- Systolic blood pressure, diastolic blood pressure between 0 and 300 mmHg
- The respiratory rate is between 0 and 100 breaths per minute.

- The pulse rate is between 0 and 300 beats per minute.
- The body temperature is between 0 and 50 ° C.
- To determine the range of the plausible body weight and height, we divided the data into three groups: patients younger than 12 months, younger than 20 years, and older than 20 years. We manually calculated the mean of values and ± 3 standard deviations for each group.

Laboratory test results Laboratory test results have variables such as the date of examination, code of order and examination, and results. In total, there are 8,088 examination codes, which are related to clinical pathology and nuclear medicine, with approximately 1.1 billion records associated with 748,474 patients. However, there are many overlapping records, because different prescription codes can be used for the same test and result. Since the examination result is more important than the path of the prescription, duplicate data were removed based on examination results. Moreover, each examination has a process in which a person enters the result directly, potentially introducing human errors, so secondary data cleansing was performed considering the types of value (integer, float, categorical, etc.) that should be the result of each examination. As a result, a total of 480 million records with 1,335 examination codes are included in the laboratory test results.

Patient history questionnaire At the time of admission, the health-related history is obtained, including information such as details of hospitalization, vital signs, personal and family medical history, past diagnoses, clinical symptoms, and lifestyle. Information on smoking, including status, period, amount per day, quit-smoking period, and quit-smoking education training status, was extracted from the questionnaire data. Based on a review of pack-year distribution in the patient population, the range was found to be generally between 0 and 500, which suggests that the data can be considered relatively reliable. With regard to the disease history, we identified some diseases that could be considered as complications of CVDs, such as diabetes, hypertension, tuberculosis, and hepatitis. We obtained patient history data, including marital status, exercise, religion, and education, along with smoking history.

Unstructured data

The major CVD-related examinations are echocardiography, Holter monitoring, ECG, thallium SPECT, and coronary artery CT. The majority of results from these tests are readings formatted as free-text, recorded as a mixture of Korean, English, numerical values, and special characters. Despite the presence of numerous errors and nulls introduced by the fact that most of the entries were handwritten, there are significant variables associated with CVDs. Therefore, the process of data formalization through NLP is essential.

We performed NLP on the readings of major digital examinations related to CVDs, deriving variables with high clinical importance that can be directly applied to AI models. The basic method of NLP applied to unstructured data can be described in three steps: First, we created a meta-table consisting of the main variables and conditions of extraction by the clinician. Second, we divided the readings into three frames: text, tabular, and others, and defined the extraction rules for each frame. We took into consideration the structure of the original data and the location of variables set in the meta-table and defined rules using a variety of operators and regular expressions. Third, the new tables were built by extracting the keywords and features from the original data. The values of keywords were based on rules defined in the previous step. This approach was used to process the six-minute walk test, pulmonary function test, cardiac rehabilitation, pediatric echocardiography, and treadmill tests. A simplified NLP flow chart is depicted in Figure 3.

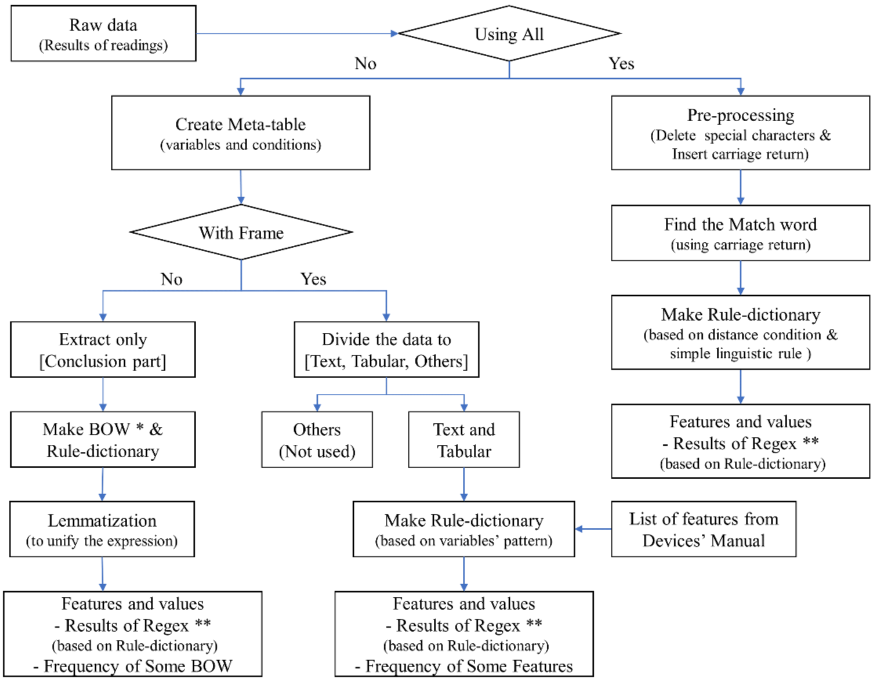


Figure 3 The flow chart of natural language processing. *Bag-of-words (BOW); **Regular expression (Regex)

Additionally, information extracted from PACS contains records of CVD-related tests performed in over 96% of patients, including outcomes of cardiac examinations, imaging, interventional procedures, and arrhythmia assessments. We performed NLP to extract the examination codes to help classify and layer data in other tables. The specific methods used for echocardiography, Holter monitoring, thallium SPECT, and coronary CT are as follows:

Echocardiography Compared with other digital tests, echocardiography is the most common test performed in patients with CVDs, with a total of 538,630 patients (71% of the entire CardioNet population) undergoing this test. Since the results of echocardiography are sentences in free-text form without a frame, this information was processed with specific rules for extracting values.

First, the "Conclusion" part of the reading, which includes a summary of important values of test results and the clinicians' interpretation, is set to the extraction range. As primary verification, we investigated all words in the "Conclusion" section and corrected typographical and grammatical errors. Second, we tokenized the words and created the bag-of-words (BOW) that contain the keywords and their frequencies [12].

Approximately 9,000 keywords were identified, appearing a total of 3.4 million times. We subsequently created a rule-dictionary to consistently replace errors with the correct words, since it is not always possible to scrutinize all the data. Third, as the secondary verification, the lemmatization, which is a type of normalization, was conducted. English words differ in terms of morphology depending on parts of speech or tense, so we extracted the stem of the word (the core part containing the meaning) and unified the expression, allowing it to be recognized as the same keyword. Since the meaning of the word may vary depending on context and/or affix, we modified the keywords once more following a full investigation. About 9,000 keywords were reduced to about 2,500. Additionally, because of bias and range of pre-knowledge, cross-checking between the engineer and clinician was repeated twice to improve the completeness and accuracy of the rule dictionary. Based on the clinician's opinion and review of CVD-related research, we selected approximately 100 meaningful features among the extracted keywords and created an echocardiography table by defining patterns and extracting them into binary or continuous values.

Holter monitoring Holter monitoring is the examination performed using electrodes and a recording device to track the heart's rhythm for a period of 24 to 72 hours. Although a total of 61,771 patients had undergone Holter monitoring (less than the number of patients subjected to other tests), Holter monitoring test is essential for patients with irregular cardiac rhythm. Holter monitoring tends to be more regular than other examinations because the readings are automatically generated by the test equipment. Since AMC is using the General Electric equipment, we were able to obtain a list of variables that appear in the records from the equipment manual. Meaningful keywords were subsequently selected through interpretation by the clinicians, with appearance and frequency expressed using an approach similar to that used with echocardiography data.

Thallium SPECT Thallium SPECT is the examination used to diagnose coronary diseases and to verify the survival of the myocardium. A total of 156,615 patients underwent thallium SPECT at AMC. In previous research, the values of summed stress score, summed rest score, and summed difference score derived from the thallium SPECT equipment were used. However, these values often do not match those in the patient records. We, therefore, conducted NLP by considering the perspectives of the clinical field. As with other modalities described above, rules were defined to find the required information, including the output from the device, and variables and values were extracted to create the table. Specifically, the position and degree at the time of stress were extracted from the readings, allowing us to deduce the position and extent of the cardiac problem. According to the patient's condition, up to eight disorders were identified.

Coronary artery CT Coronary artery CT is another examination performed to diagnose CVDs and evaluate the presence of various cardiac conditions. A total of 79,046 patients at AMC underwent coronary CT. We modified the NLP pipeline to further increase the coverage of data. The key concept underlying the NLP method is the analysis of the text based on linguistic rules, focusing on keywords that are to be extracted. This method does not need to divide the structure of the readings into individual forms and can be performed for all sentences. We identified words on the same line by assigning a certain distance condition to the "carriage return," as well as by using a "match word." The keywords corresponding to stenosis degree and plague for each segment existing in the reading text were extracted. We completed the extraction of additional information, such as the type and patency status of vascular treatments in patients who had undergone stent graft surgery and details about the native coronary artery. After processing, we validated the data that had been structured and processed with input from clinicians and evaluated the reliability of the CardioNet data. We constructed multiple scenarios based on clinicians' practical experiences, sampled the data, and visualized the frequency, data types, and other parameters.

Standardization

Securing interoperability is essential for collaborative research with other institutions and hospitals in CVDs, as well as other diseases. We carried out the standardization process on terms and codes used in the hospitals to build the CardioNet that corresponds to the CDM form. However, term mapping is not possible without practical familiarity with the usage of the code and terms in the actual clinical field, as well as with insight into the unique information used in the institution. It is necessary to seek input from highly experienced specialists in fields such as diagnosis, surgery, laboratory tests, pathology, imaging, medication, blood transfusions, and materials. Consequently, we received counsel from the AMC medical data team and modified the codes and terms into standard terms with the same meaning.

Because there are currently no common codes for digital examination in CVDs, we created independent variables based on meaningful clinician comments, which will be converted once digital test codes are defined in the CDM (e.g., in case of echocardiography, there are 97 variables and explanation), Specifically, OMOP-CDM-based local code mapping was performed because of the differences in the code systems used in each hospital [13]. The codes of diagnosis, operation, image pathology, blood transfusion, and procedure and materials were mapped based on SNOMED-CT, laboratory test results were based on LOINC, medication was based on RxNorm. The bacterial code did not proceed in accordance with the OMOP-CDM standards [14,15,16]. The mapping ratio of AMC's codes based on CDM is shown in Table 2. The ratio is almost 90% or more, so it can be used immediately when expanding to other topics of study.

Table 2 The mapping ratio of AMC's code
SNOMED-CT: systematized nomenclature of medicine-clinical terms LOINC: logical observation identifiers names and codes
RxNorm: a standardized nomenclature for clinical drugs produced by the U.S. National Library of Medicine (NLM)
* Except for bacterial code

Class	Total codes (*N*)	Mapped codes (*N*)	Mapping ratio (%)	Remark
Diagnosis	10,728	10,708	99.81	SNOMED-CT
Surgery	1,554	1,544	99.36	SNOMED-CT
Laboratory test	705	599	84.96	LOINC *
Image pathology	247	245	99.19	SNOMED-CT
Medication	4,631	4,600	99.33	RxNorm
Blood transfusion	23	23	100	SNOMED-CT
Procedures and Materials	386	382	98.96	SNOMED-CT

Using the processed and standardized data of 748,474 patients, we constructed the CardioNet schema based on the hospital EHR structure. We created a descriptive table for 27 tables sorted by category, which is intended to facilitate database usage for clinicians and engineers, allowing them to easily access and understand the data. A descriptive table is a dictionary of data variables, complementing the dictionary of word definitions and variables in the CardioNet and explaining the anatomy, physiology, and pathology of the heart. Additionally, it displays CVD-related variables, their meaning, and clinical utility. Since most variables in the table have abbreviated values (code of orders, diagnosis, examination, etc.), we created a code master table and linked these values. This makes it easy to find the meaning of words and abbreviations.

We continuously validated the data during pre-processing to further ensure the reliability of the CardioNet and verify the processed data. Furthermore, we constructed scenarios based on the clinicians' practical experiences, sampled the data, and visualized their frequency and types.

CardioNet built

All data tables (except demographics and human-derived materials) have a KEY column concatenating PAID and INNO. The demographics table is the central table, consisting of 748,474 patients. The second central table is the visitation table with 743,332 patients. Demographic and human-derived material table without INNO is connected to the visit table by PAID, while all other tables are linked by KEY. Figure 4 describes the entity-relationship diagram (ERD) of CardioNet, with the full form of abbreviations in the ERD listed below.

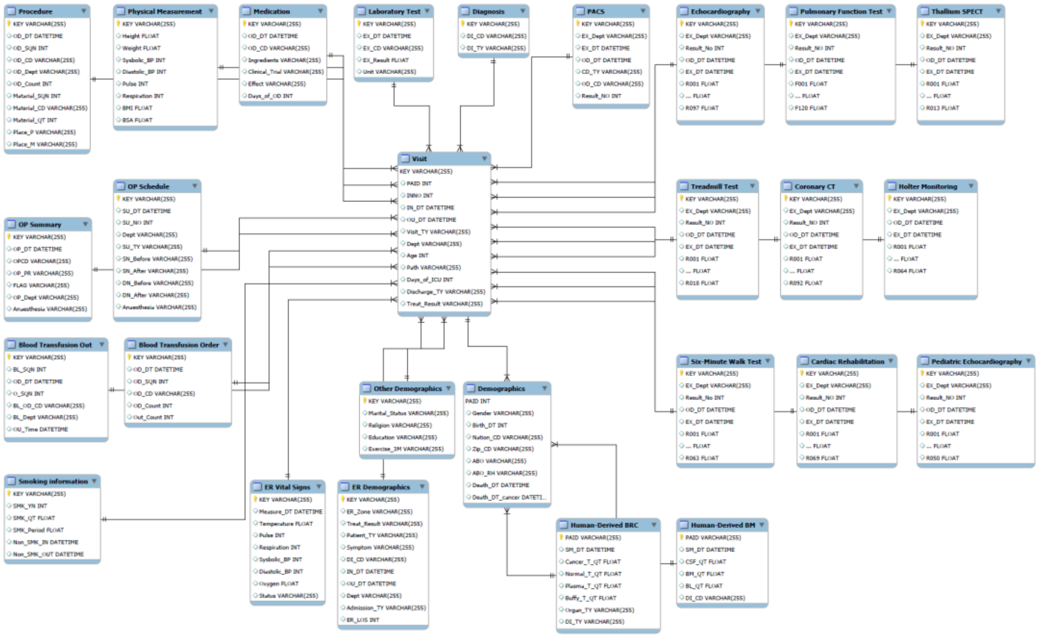


Figure 4 ERD of CardioNet; ERD: entity-relationship diagram

Results

A total of 74.8 million patients visited AMC or UUH for CVDs between January 2000 and December 2016. CardioNet is a comprehensive database intended to support the development of predictive models of CVDs and future multi-center convergent research. It comprises information that can be extracted from EHRs and has undergone structuralization and standardization by the processing of the readings of CVD-related digital tests by NLP. CardioNet contains a total of 27 tables, a code-master table, and a descriptive table.

Summary of CardioNet

Table 3 summarizes the tables of CardioNet, providing a description of individual tables, number of features, and the number of records and patients included from each hospital.

Table 3 Summary of CardioNet

Description	Features (N)	Number of records		Number of patients	
		AMC	UUH	AMC	UUH
Demographics	9	572,811	175,663	572,811	175,663
Demographics (ER)	20	502,055	171,489	214,393	72,423
Vital signs (ER)	13	1,865,348	-	185,447	-
Physical measurement	14	46,768,559	5,485,196	511,061	130,361
Visits	23	18,967,703	8,935,764	571,163	172,169
Diagnosis	13	28,328,713	8,089,345	553,031	174,403
Schedule of operation	12	434,085	-	245,159	-
Summary of operation	14	3,404,439	88,760	348,939	52,852
Six-minute walk test	74	32,158	1,210	8,871	665
Coronary artery CT	97	97,585	-	79,046	-
Thallium SPECT	26	198,711	-	156,615	-
Echocardiograp hy	112	726,187	178,386	428,004	110,626

Holter monitoring test	75	66,366	21,035	46,636	15,135
Pulmonary function test	135	4,634,091	63,593	265,817	38,933
PACS	12	12,410,683	4,490,786	551,280	169,801
Pediatric echocardiography	63	4,017	-	1,720	-
Cardiac rehabilitation	80	2,912	-	1,990	-
Treadmill test	29	110,094	31,741	68,203	25,979
Laboratory test	7	344,908,032	143,847,546	489,278	175,663
Medication	26	129,804,022	57,639,868	500,444	162,750
Procedures and materials	21	105,739,326	13,201,735	417,407	136,128
Order of blood transfusion	10	1,090,115	219,804	192,169	43,814
Result of blood transfusion	11	2,764,232	625,574	100,215	28,621
Human-derived materials	13	46,760	-	43,412	-
Human-derived bonemarrow	13	5,757	-	2,983	-
Patient history	10	673,143	-	307,681	-
Smoking information	12	608,441	-	280,492	-

Demographics

A total of 572,811 patients visited AMC and 175,663 patients visited UUH. Table 4 depicts the demographics of the two hospitals, including the physical measurements and the number of patients who have undergone CVD-related digital examinations. Approximately 45% of patients were women, with an average age of 55.78 years at the time of the initial encounter. Body measurements, such as weight and height, and vital signs, such as blood pressure, are not consistently performed at each visit. As shown in Table 4, the average valid value was calculated for each patient (i.e., 563,131 patients and 543,792 patients have valid blood pressure and BMI values calculated). Blood pressure data are available for 75.23% of patients, with the average systolic blood pressure determined to be higher than 120 mmHg and therefore above the normal range, while average diastolic blood pressure was below 80 mmHg.

The WHO Asia-Pacific region and the Korean Obesity Association standards deem individuals as overweight when their BMI is 23 kg/m22 or higher and obese when BMI is 25 kg/m22 or higher. According to these criteria, the average BMI values show the majority of the patients to be overweight [17]. Additionally, 63.89% of patients visited the department of cardiology or thoracic surgery more than once, with 39.17% patients registering more than three visits. Patients with CVD-related diseases continued to visit the hospital. In processing the digital medical tests, duplicates were removed for each patient, and the number of cases examined more than once was determined. This analysis demonstrated that 71.96% of the patients underwent echocardiography.

Table 1 Demographics
**N* of Blood Pressure: AMC = 461693, UUH = 101438
***N* of BMI: AMC = 457621, UUH = 77171
****N* of Visits total: AMC = 571163, UUH = 172169

	AMC **(*N* = 572,811)**	**UUH** **(*N* = 175,663)**	**Total** **(*N* = 748,474)**
Gender ([F,M])	[257160, 315651]	[79988, 95675]	[337148, 411315]
Age (Year)	56.32 ± 14.72	52.11 ± 18.09	55.78 ± 15.20
Systolic blood pressure (mmHg)	123.06 ± 12.61	129.05 ± 13.38	124.14 ± 12.95
Diastolic blood pressure (mmHg)	74.29 ± 7.94	75.96 ± 9.07	74.59 ± 8.18
BMI (kg/m^2) **	24.11 ± 3.50	24.04 ± 3.55	24.100 ± 3.513
CV/CS Encounter(*N*) ***			
0	250,160	14,925	265,085
1	68,037	19,489	87,526
2	78,406	19,101	97,507
≥ 3	174,560	118,654	293,214
Test (N (%))			
Echocardiography	428,004 (74.71%)	110,626 (62.97%)	538,630 (71.96%)
Pulmonary function	265,817 (46.40%)	38,933 (22.16%)	304,750 (40.71%)
Thallium SPECT	156,615 (27.34%)	-	156,615 (20.92%)
Treadmill	68,203 (11.90%)	25,979 (14.78%)	94,182 (12.58%)
CT	79,064 (13.80%)	-	79,064 (10.56%)

Holter monitoring	46,636 (8.14%)	15,135 (8.61%)	61,771 (8.25%)
Six-minute walk test	8,871 (1.54%)	665 (0.37%)	9,536 (1.27%)
Cardiac rehabilitation	1,990 (0.34%)	-	1,990 (0.26%)
Pediatric echocardiography	1,720 (0.30%)	-	1,720 (0.22%)

Visits

The total number of visits to the Departments of Cardiology or Thoracic Surgery is presented in Table 5, which shows the number of rows related to visits to each department. A total of 428,247 patients visited the departments more than once, accounting for 57.21% of the total. The total number of visits by these patients to both hospitals is approximately 4.69 million, accounting for 16.82% of the total number of visits (27.9 million). As shown in Table 5, the average age of these patients was 58.8 years, which is 3.01 years higher than the average age of the entire patient population. Outpatient visits comprised 92% of all visits, inpatient visits accounted for 4.86%, and approximately 3% corresponded to ER visits (with only AMC ER data considered).

Table 2 Number of visits to the departments of Cardiology or Thoracic surgery

	AMC **(*N* = 321,003)**	**UUH** **(*N* = 157,244)**	**Total** **(*N* = 478,247)**
Age (Year)	59.85 ± 13.21	57.28 ± 15.00	58.80 ± 14.03
Outpatients	2,548,245	1,854,432	4,402,677
Inpatients	134,846	71,012	205,858
ER	86,429	-	86,429

Diagnosis

Table 6 describes the number of patients diagnosed with nine major CVDs and complications. Because a single patient can be diagnosed with multiple diseases and diagnosis records are taken at each visit (Table 6), duplicates were removed and each patient's unique diagnostic names were counted. A total of 445,787 patients (59.55%) were identified with major CVDs. Hypertension was diagnosed most frequently (31.79% of the entire CardioNet population), followed by throat and chest pain, diabetes mellitus (including types 1 and 2, malnutrition-related, unspecified, etc.), angina pectoris, chronic ischemic heart diseases, cerebral infarction, heart failure, acute myocardial infarction, and cardiac arrest. Considering that the average of the number of diseases for each patient diagnosed with the major CVDs was 1.82 (standard deviation 1.2) and that there are patients with up to nine CVDs, a number of patients were found to exhibit comorbidities and complications.

Table 3 The number of patients with major diseases

Diagnosis	AMC (*N* = 357,910)	UUH (*N* = 87,877)	Total (*N* = 445,787)
Hypertension	200,109 (55.91%)	37,886 (43.11%)	237,995 (53.38%)
Pain in throat and chest	142,567 (39.83%)	38,690 (44.02%)	181,257 (40.66%)
Diabetes mellitus	112,381 (31.39%)	30,236 (34.40%)	142,617 (31.99%)
Angina pectoris	61,789 (17.26%)	9,694 (11.03%)	71,483 (16.03%)
Ischaemic heart disease	47,836 (13.36%)	5,847 (6.65%)	53,683 (12.04%)
Cerebral infarction	24,752 (6.91%)	8,958 (10.19%)	33,710 (7.56%)
Heart failure	15,345 (4.28%)	4,825 (5.49%)	20,170 (4.52%)
Acute myocardial infarction	10,543 (2.94%)	3,853 (4.38%)	14,396 (3.22%)
Cardiac arrest	1,213 (0.003%)	1,196 (0.013%)	2,409 (0.005%)

Laboratory results

The laboratory test table contains the results of 1,335 diagnostic tests, with a total of 480 million rows of data from 664,941 patients (89% of the total population). In some cases, a single patient may undergo multiple tests during the day or the same tests several times a day. Table 7 presents the percentage of patients who have undergone each laboratory test sorted by the number of patients seen at AMC. Excluding duplicate entries, results of the following laboratory tests were found for at least half of all patients: creatinine, cholesterol, alanine transaminase, aspartate transaminase, bilirubin (total), albumin, protein, glucose, alkaline phosphatase, hemoglobin, platelets, calcium, uric acid, potassium, sodium, blood urea nitrogen, chloride, CO22(total), and phosphorus. Results of other tests were available for at least 15% of the patients, including triglycerides, high density lipoprotein-cholesterol, low density lipoprotein-cholesterol, C-reactive protein (quantity), erythrocyte sedimentation rate, hemoglobin A1c, creatine kinase, troponin-1, and high-sensitivity C-reactive protein.

Table 4 Percentage of patients with laboratory results

Laboratory test	AMC (%)	UUH (%)	Total (%)
Creatinine	83.64	91.24	85.42
Cholesterol	83.6	93.25	85.86
ALT	83.53	92.21	85.57
AST	83.53	92.26	85.58

Bilirubin (total)	83.03	91.29	84.97
Albumin	83.02	91.34	84.97
Protein	83.01	91.3	84.96
Glucose	83	90.47	84.75
ALP	82.97	91.25	84.91
Hb	82.88	92.51	85.14
Platelet	82.88	92.48	85.13
Calcium	82.77	87.37	83.84
Uric acid	82.68	89.44	84.27
Potassium	80.54	87.91	82.27
Sodium	80.51	87.95	82.25
BUN	75.36	91.08	79.05
Chloride	65.94	87.42	70.98
CO2 (total)	65.89	83.48	70.02
Phosphorus	62.4	87.4	68.27
Triglyceride	53.61	37.55	49.84
HDL-Cholesterol	52.96	37.17	49.26
LDL-Cholesterol	44.09	31.45	41.12
CRP (quantity)	42.94	67.49	48.71
ESR	42.79	42.7	42.77
Hb A1c	38.66	32	37.09
CK	27.76	43.13	31.37
Troponin-I	25.48	11.62	22.23
hsCRP	17.64	14.67	16.94

Echocardiography

Echocardiography accounted for the largest number of medical digital examinations undergone by patients (71.96%). Echocardiography readings were converted into 112 variables, with the basic statistics of 19 major variables reflecting clinicians’ opinion shown in Table 8. The descriptive table includes the meaning and clinical interpretation of each variable. As an example, the left ventricle (LV) dimension in systole refers to the inner diameter of the LV measured during systole, with the expansion of the LV indicated when the value in a male patient exceeds 42 mm.

Table 5 Prime features in echocardiography
LV: Left Ventricular, LVESD: LV dimension in end-systole, LVEDD: LV di- mension in end-diastole, TR: Tricuspid regurgitation, LVPWES: LV poste- rior wall thickness in end-systole, LVPWED: LV posterior wall thickness in end-diastole, LVIVSES: LV interventricular septum thickness in end-systole, LVIVSED: LV interventricular septum thickness in end-diastole, LAd: Left atrial diameter, LVESV: LV volume in end-systole, LVEDV: LV volume in end-diastole, LVEF: LV ejection fraction

Description	**AMC (*N*=428,004)**	**UUH (*N*=110,626)**
LVESD	30.25 ± 6.43	30.30 ± 5.76
LVEDD	47.77 ± 8.74	47.20 ± 5.57
LVPWES	13.93 ± 2.90	14.32 ± 2.25
LVPWED	9.02 ± 1.87	9.25 ± 1.55
LVIVSES	13.14 ± 2.88	13.42 ± 2.26
LVIVSED	9.09 ± 2.01	9.52 ± 1.74
LAd	37.39 ± 8.72	36.04 ± 6.23
LVESV	35.99 ± 18.94	36.83 ± 23.61
LVEDV	88.97 ± 32.39	80.82 ± 33.07
E/A ratio	0.93 ± 0.52	0.44 ± 0.54
E/E ratio	8.52 ± 7.18	9.64 ± 3.63
LVEF	58.83 ± 11.93	62.18 ± 8.47
LV mass	163.78 ± 57.80	149.31 ± 45.74

Discussions

The EHRs are easily accessible in hospitals and contain important clinical information. EHR data were found to be less useful in AI studies, compared to data from imaging modalities such as CT or MRI [5-9]. We summarized the insights and expectations of building CardioNet along with its limitations.

First, in the AMC, there is cloud environment, which is an infrastructure for research with researchers inside and outside the hospital. This cloud system aims to implement a shared database and collaboration for multi-center medical AI research. Also, it is a hybrid cloud that can use a public cloud (e.g., AWS, MS Azure) based on a private cloud. Therefore, those who want to do research with CardioNet, need to register as a joint research team to the IRB according to hospital policy. Currently using CardioNet, researchers outside the AMC who have access to cloud are conducting various studies.

Second, it is difficult to directly use EHR data extracted from a hospital system to generate meaningful results with AI research. This reflects the fact that EHRs include numerous unstructured free-text entries. These free-text readings contain important clinical insights made by clinicians as part of patients' treatment. In setting up AI research, it is necessary to decide how to handle free-text readings. The simplest way is to process the free-text notes as sentences and infer the meaning between words, as in the NLP field study [18,19,20]. This approach is in line with a number of existing studies that derive useful meaning by analyzing the medical articles themselves by NLP [21]. However, the biggest problem with this method is that English in free-text readings differs between clinicians, hospitals, and countries, unlike the English used in scientific communication within research articles. In different countries, these readings can be mixtures of the official national language and English. Therefore, the use of free-text readings in AI research requires significant support from clinicians. Clinicians alone can understand free-text readings, with specialized expertise in relevant topic needed for accurate interpretation.

We worked with clinicians to find valid patterns of readings, and performed rule-based NLP to convert results into variables and values. We considered various ways to apply NLP to free-text readings, but in most cases required a lot of human labor. Therefore, we are working on more efficient automatic NLP research that can be applied regardless of locality.

Also, in the process of performing NLP, we created a descriptive table (i.e., dictionary of CardioNet) that describes clinically valid variables. Although several cardiologists at AMC and UUH participated and reviewed this study, some extracted variables can be subjective compared to the numerical values. Therefore, we plan to strengthen the descriptive table by sharing it with the Korean Cardiology Association to collect the opinions of other cardiologists.

Third, in building CardioNet, we were assisted by numerous cardiologists, with most input involving standardization of the free-text readings. This is a common issue in building an EHR-based database for AI research in all types of diseases, including CVDs. The current CDM for multi-center clinical research is not suitable for AI research. CDM is a good example of the standardization of deidentified patient data, with a number of hospitals building CDM-based datasets for clinical research. However, unlike clinical studies that focus on patient events, the important feature of EHR in AI research is the time-series data. The performance of AI-based predictive research based on EHRs is determined by how well the changes in the patient's state over time are embedded in the training data. Therefore, it is necessary to develop a CDM for clinical AI research that standardizes patient events over time.

Finally, we are preparing a number of AI-based studies including automatic NLP as future work using CardioNet. Recently, deep learning technology in medical images is developing with results that exceed experts, but it is also necessary to apply AI to data linking raw images and EHR. Since the honest broker in the de-identification process has a key, the CardioNet and image and signal raw data in PACS and hospital local databases can be connected. This connection is essential for the realization of patient-specific medical care, and we are carrying out related research. As a result, we are in the process of developing and characterizing an AI model that can perform CVDs risk prediction according to the characteristics of CVDs, where prevention is more important than diagnosis.

Conclusion

We established the comprehensive database specialized in CVDs. We are actively supporting multi-center research, which may require further data processing, depending on the subject of the study. CardioNet will serve as the fundamental database for future CVD-related research projects.

Chapter 2. Machine Learning–Based Hospital Discharge Prediction for Patients with Cardiovascular Diseases: Development and Usability Study

Introduction

Background

The use of human and physical resources, which are both costly and scarce, is essential for the efficient operation of hospital processes. Hospitals are required to manage different kinds of resources, like managing the schedules of the medical team and staff, bed management, and clinical pathways to improve overall management efficiency [22]. Effective resource management in hospitals can improve the quality of medical services by reducing the labor-intensive burden on staff, decreasing inpatient waiting time, and securing optimal treatment time [23].

Bed management is a form of hospital resource management. Currently, in most hospitals, clinicians manually check a patient's condition to decide whether to continue their hospitalization or discharge them [24]. On the basis of this decision, the medical team and staff identify the bed capacity available in the near future and schedule the patient's reservation. In addition, the number of patients hospitalized for a variety of chronic and acute illnesses, such as cardiovascular diseases (CVDs) [25], has been steadily increasing, and their insufficient treatment can lead to readmissions or complications. However, a stay in the hospital longer than the optimal treatment time hinders effective bed management. Thus, it is important to accurately predict the patient's hospitalization period and make judicious decisions about their discharge.

Many studies have focused on the efficiency of hospital resources, and most of them presented algorithms or models for improving bed management. Bachouch et al [26] investigated hospital bed planning and proposed the integer linear program to solve the optimization problem. They illustrated the simulated bed occupancy schedule. Troy et al [27] studied the simulation of beds for surgery patients using the Monte Carlo simulation to determine the intensive care unit (ICU) capacity. Particularly, the predicted length of stay (LOS) is one necessary piece of information for bed management, and there are many studies predicting the LOS based on electronic health records (EHRs) [28-30].

Moreover, authors have used machine learning (ML)–based models to predict the LOS [28-30], prolonged hospitalization, and unplanned readmission [31] and to find biomarkers for critical diseases [32]. Recently, there have been many studies on interpretable or explainable artificial intelligence (XAI) [33]. One XAI study [34] developed a model to predict acute illness and provide results and interpretation. Compared with EHRs, studies employing computer vision algorithms such as convolutional neural networks are more actively pursued because these models can directly visualize significant parts of an image [35,36]. Thus, we developed an ML-based predictive model to provide the daily discharge probability and individual explainer visualizing significant features of each patient to support bed management.

Objectives

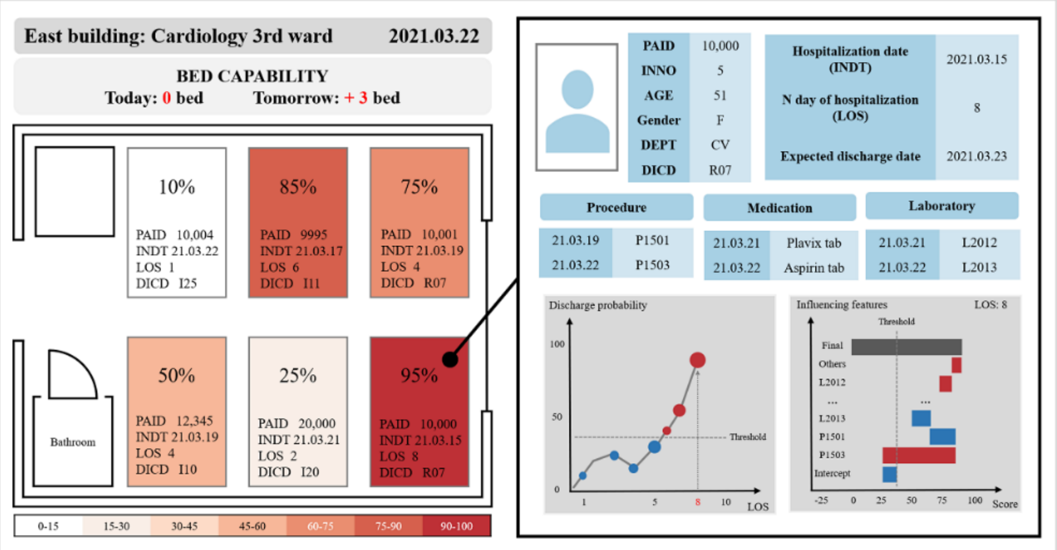
The main contributions of this study can be summarized in the following steps: first, we developed an ML-based predictive model to predict the discharge probability daily within 3 days for each patient with CVDs and to acquire the individual LOS. Patients with chronic and acute diseases, including CVDs, have high hospitalization and readmission rates and greater complications [37]. There are alternatives to transfer those who need urgent care or hospitalization to another hospital to address delays. However, it could be causing other serious problems, hospitals should continuously identify methods to reduce waiting time, and efficient bed management can be considered as one of them.

In addition, because of the diversity of diseases, it may be more advantageous to find common risk factors and implement bed management for specific departments or diseases (ie, clustered specific wards), and then expand it further to the hospital level. Therefore, we developed an ML-based model to determine the bed capacity that would be available in the near future and find risk factors by predicting the discharge of patients hospitalized with CVDs [38]. By providing persuasive discharge information such as expected individual discharge date and risk factors related to CVDs, it is possible, in practice, to assist in precise bed management, which is otherwise done manually by the medical team.

Second, we assessed the outcome of the prediction and provided the individual explainer to describe the primary risk factors of inpatients for patient-specific care. Even if patients have the same diseases and common variables represent the diseases, each patient has different characteristics, history, circumstances, and treatments. Therefore, it is also necessary to identify and monitor the unique, individual variables for each patient. In this study, our ML-based predictive model's outcomes include not only information on daily patient discharge but also the contributions of features such as feature importance. Furthermore, we visualized the day-by-day discharge probability of each patient and the features that influenced individual patients during the hospitalization. This explainer can guide the medical team and patients to produce reasonable evidence on the ML-based model's outcomes and helps them understand the conditions in detail and prepare in advance for treatment. Such individual analysis can focus on each patient, and the meaningful features identified can be used in other studies as a basis for preidentifying variables affecting hospitalization.

Third, this study could help manage bed scheduling efficiently and detect long-term inpatients in advance. Bed management refers to the process of identifying patients who are most likely to be discharged, confirming the number of available beds, and allocating beds to patients waiting for admission after reservation. As this process is complicated and usually carried out manually, we aimed to support it by providing the estimated LOS and probability of discharge returned by the model and by identifying the capacity of beds that would be available in the near future. In addition, it is possible to detect not only patients with a high probability of discharge but also patients with a consistently low probability of discharge. In other words, it helps discover and analyze the causes of long-term hospitalization of high-risk patients and provides this information to their management team.

To summarize, we developed an ML-based model to predict whether hospitalized patients with CVDs would be discharged within 3 days. On the basis of this model, we proposed an individual explainer; the simulations of bed management are depicted in Figure 5, including the probability of discharge and influenced features such as demographics, prescribed medications, and treatments. Our model can improve the efficient use of hospital resources and enhance the quality of medical services.



PAID	10,000
INNO	5
AGE	51
Gender	F
DEPT	CV
DICD	R07

Hospitalization date (INDT)	2021.03.15
N day of hospitalization (LOS)	8
Expected discharge date	2021.03.23

Procedure		Medication		Laboratory	
21.03.19	P1501	21.03.21	Plavix tab	21.03.21	L2012
21.03.22	P1503	21.03.22	Aspirin tab	21.03.22	L2013

Figure 5 Visualized simulation of discharge prediction for machine learning–based bed management. DEPT: department; DICD: diagnostic code; INDT: the date of visitation or admission; INNO: the patient's encounter number; LOS: length of stay; PAID: the patient's identification.

Methods

Overview

Figure 6 describes the overall flow of the prediction method employed in this study. We set up the cohort criteria and processed the data to create suitable data sets. Subsequently, we trained the ML-based predictive models and evaluated them to find an elaborate model. Finally, we predicted the discharge probability within 3 days and explained the model's outcomes by identifying, quantifying, and visualizing its features.

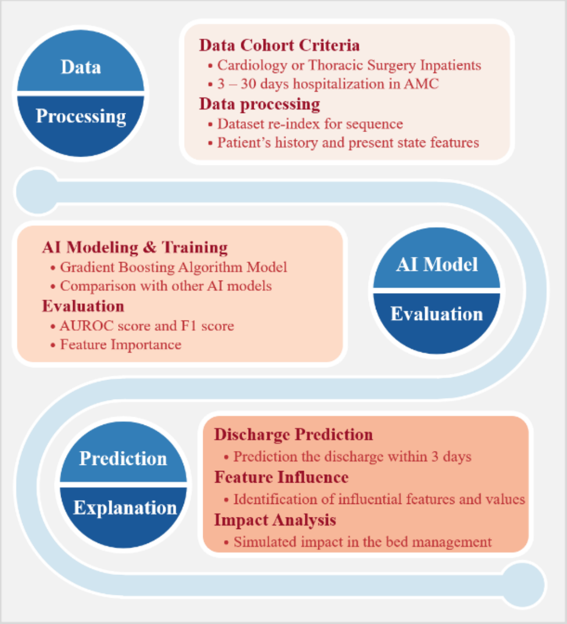

Data
Processing
Data Cohort Criteria
• Cardiology or Thoracic Surgery Inpatients
• 3 – 30 days hospitalization in AMC
Data processing
• Dataset re-index for sequence
• Patient's history and present state features
AI Modeling & Training
• Gradient Boosting Algorithm Model
• Comparison with other AI models
Evaluation
• AUROC score and F1 score
• Feature Importance
AI Model
Evaluation
Prediction
Explanation
Discharge Prediction
• Prediction the discharge within 3 days
Feature Influence
• Identification of influential features and values
Impact Analysis
• Simulated impact in the bed management

Figure 6 Overall flow of the prediction method for discharge within 3 days. AI: artificial intelligence; AMC: Asan Medical Center; AUROC: area under the receiver operating characteristic.

Data Acquisition

Data were extracted from CardioNet [39] (Textbox 1), a manually curated EHR database specialized in CVDs. CardioNet consists of data from 572,811 patients who had visited Asan Medical Center (AMC) with CVDs between January 1, 2000, and December 31, 2016. The AMC institutional review board approved the collection of CardioNet data and waived informed consent. CardioNet contains 27 tables on topics such as visitation, demographics, diagnosis, medication, and laboratory examination. Most tables have common variables including patient identification (PAID), patient encounter number (INNO), the date of visitation or admission (INDT), and the date of discharge (OUDT). The KEY column, which concatenates the PAID and INNO columns, can connect the visitation table to other tables. Using the KEY column, we extracted the variables in each table to be analyzed.

Textbox 1 Data extracted from CardioNet.

Data extracted from CardioNet

- Visit table: patient identification, patient encounter number, KEY, date of visitation or admission, date of discharge, type of visit, medical department, and duration of stay in the intensive care unit (ICU).
    - acute care unit, coronary care unit, cardiac surgery ICU, medical ICU, neonatal ICU, neurological ICU, neurosurgical ICU, pediatric ICU, and surgical ICU.
- Diagnosis table: International Classification of Diseases-10th edition code of diagnosis.
- Laboratory test result table: date and code of pathology examination, and the result of the examination.
- Physical information table: patient’s age, height, weight, systolic and diastolic blood pressures, respiratory rate, pulse rate, BMI, body surface area, and date of measurements.
- Medication table: date and code of prescription.
- Procedure table: date and code of order.
- Operation table: date and code of surgery or treatment.
- Picture archiving and communication system table: date and code of order.
- Transfusion order table: date and code of order.

From the 572,811 patients in CardioNet, we obtained 84,251 records of 63,261 anonymous patients hospitalized in the departments of cardiology or thoracic surgery. Furthermore, to develop a practical and usable model, we focused on predicting discharge within 3 days and detecting long-term patients. Long-term patients, defined as those hospitalized for more than 30 days, are separately managed by the AMC. Therefore, we set the LOS between 3 and 30 days.

Data Preprocessing

Data Set Creation

In the visit table, which is the primary table of CardioNet, there are 4 main columns: PAID, INNO, INDT, OUDT, and visit-related variables. Each row represents a single hospitalization case for each inpatient. We reset the index to create a new data set with the duration between admission and discharge as date-index (eg, a row with an INDT of 2021.2.1 and an OUDT of 2021.2.10 has an LOS 10 of days; therefore, it was converted to 10 rows with 10 date-indexes). Finally, after preprocessing all values corresponding to PAID, INNO, and date-indexes of other tables, we merged and concatenated the tables to generate a new data set for model training.

Figure 7 shows the data set creation process. Each table of diagnosis, medication, laboratory test results, and physical information was used for both past and present features. The operation, procedure, and picture archiving and communication system (PACS) were used for the present features, and LOS in the ICU was used for the past features. The preprocessing of values for each table is discussed in the next section. The specific methods of feature handling are as follows:

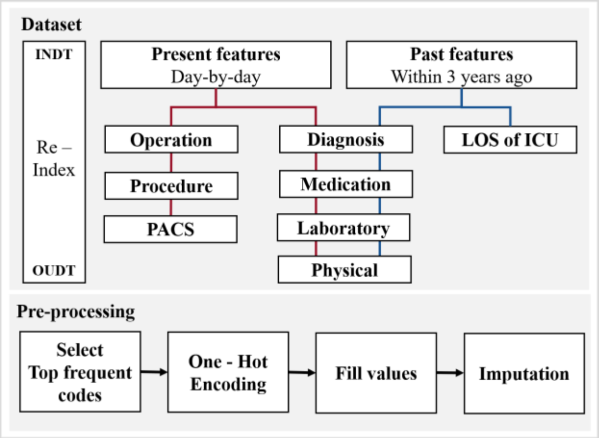


Figure 7 Data set creation process for machine learning–based model training. ICU: intensive care unit; INDT: date of visitation or admission; LOS: length of stay; OUDT: date of discharge; PACS: picture archiving and communication system.

Data-Related Features

After creating the new data set, we removed the OUDT containing future information. To distinguish and recognize the time information in date by type, we created a total of 10 date-related features. INDT and date-index were sliced into integer features such as year, month, day, and weekday. Furthermore, we created a feature that denotes whether the date-index is a holiday or not and another feature that indicates the LOS at the date-index by subtracting INDT from the date-index.

Day-by-day Present Features Related to Hospitalization

As the visit table and other tables contain only one piece of information per row, it is difficult for the ML model to learn the data all at once. Therefore, we performed one-hot encoding (OHE) of clinically important orders and codes and created them as features in the new data set. Consequently, we could access aggregated records by date for each patient.

First, in the diagnosis and operation tables, we sliced all the values of the International Classification of Diseases-10th edition codes and the operation codes at the third digit to convert them into three-digit codes because the strings from the fourth digit onward represent the subhierarchy of the three-digit codes. We arranged all the frequency values in descending order and selected the first 99 codes. We transformed the remaining codes (ie, unselected codes) into the others feature and performed the OHE on all 100 codes. The features in the form of Z_code, such as Z_DICD and Z_OPCD, refer to others in each original table. As a result, we obtained a total of 100 codes for each table (ie, diagnosis and operation table) and filled the date-index values with 1 if there were valid prescribed or ordered data and 0 otherwise. Similarly, the values of the PACS table were converted to 100 features.

Second, similar to the diagnosis table, in the medication and procedure tables, we obtained the 99 most frequent codes and others, performed the OHE, and filled the corresponding data. In the case of the transfusion table, we used all 27 codes available. We filled the values with the number of prescriptions per day or at once, considering the severity of each patient's ailment.

Third, in the laboratory test result table, the 60 most frequent examination codes, examined in more than 50% of all patients, were selected. The physical information table had only 10 codes, which were all used. We performed the OHE of values and filled them with results corresponding to each examination. If a patient had been tested several times a day, the data set was populated with the average of the results.

Past Features

We considered that the patient's anamnesis (ie, medical history) should also be included in the data set, along with the day-to-day features (described in the previous paragraph) for the ML model to learn the data deeply. When the date-index in each hospitalization started from INDT, we created some past features from the principal information of hospital visit records 3 years before INDT.

For past features, OHE was performed, and values were filled in, similar to the present features. The hospitalization periods of all ICUs in the visit table were summed up. For 100 diagnostic codes, we summed up each value if there was a record of diagnosis. For 100 medication codes, the number of prescriptions per day or at once were summed up if the record existed. Finally, recent laboratory test results and physical information within 3 years were used for a total of 70 codes. In conclusion, the data set was filled with either summed up or recent values equivalent to each feature.

Imputation

Except for the laboratory, physical, and date-related features, we replaced all the null values with zero. The value type of most of the other features was null or integer because most were calculated by frequency. In contrast, to deal with missing values in the continuous data type of the present laboratory and physical features, we first separated the data set based on the KEY. The KEY refers to a single hospitalization case of one patient; thus, separating the data set by KEY does not mix individual hospitalizations. Therefore, we filled in null values in chronological order (ie, from past to present). Subsequently, we filled in the rest of the null values in reverse chronological order (ie, from present to past) to handle those cases where results were not measured at the beginning of the admission. Using this method, it was possible to impute the null value for each hospitalization of an individual patient. Finally, to fill the values where all the features were not ordered or measured, we filled the rest of the null values with the most frequent value for each feature.

Target Criteria

The supervised learning algorithm for classification requires the label true or false to indicate the correct answer. The target criteria for true labeling in this study are depicted in Figure 8.

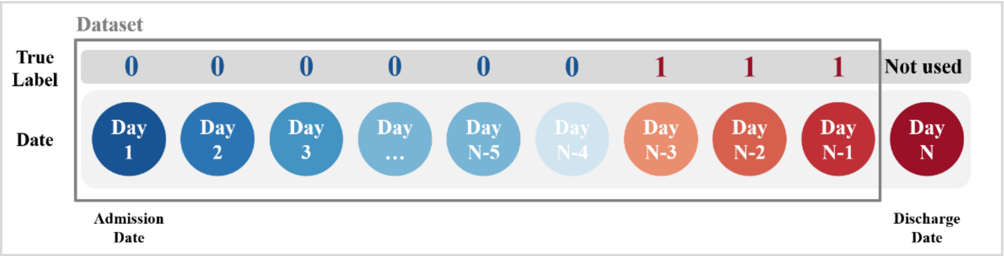


Figure 8 Target criteria to provide the true label (ie, correct answer) to machine learning–based models.

As shown in Figure 8, day 1 is INDT, day N is OUDT, and the circles represent each day of the hospitalization period. We excluded day N (ie, discharge date) from the data set because of information such as discharge procedure, which could provide the ML model with a hint. In addition, even if the accuracy of discharge prediction is higher from the discharge date to 2 days earlier, it is useful to make the prediction 3 days in advance when actually using the model. Therefore, we labeled 1 from one day before OUDT to 3 days before OUDT and labeled 0 from the INDT to 4 days before OUDT.

As a result, we transformed the diverse variables of original tables into 10 date-related features, 597 present features, and 279 past features, creating a data set of 669,667 rows with 886 features from 84,251 records of 63,261 inpatients with CVDs.

ML-Based Predictive Models

ML-Based Models

We experimented with 5 models to identify the most suitable one. We set the logistic regression [40] model as the baseline to estimate performance, and support vector machine [41,42], random forest (RF) [43], multilayer perceptron (MLP) [44], and extreme gradient boosting (XGB) [45] were selected as comparison models. We also performed hyperparameter tuning for each model through random search.

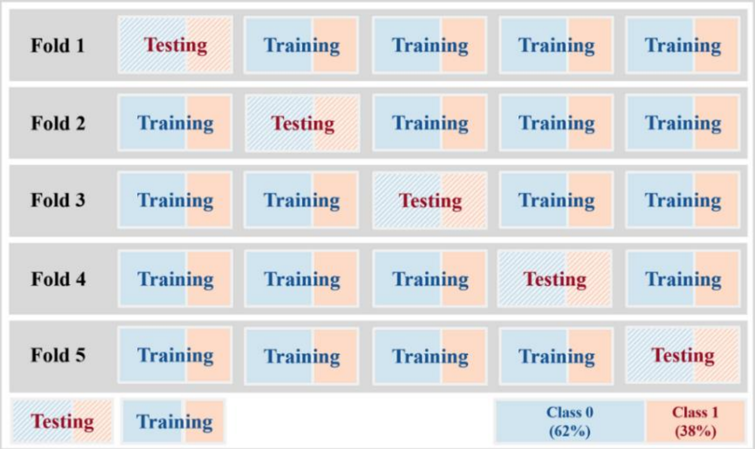
We selected XGB, which is a gradient boosting algorithm (GBM) model, as the final model. GBM is an ensemble method that combines several weak classifiers (trees). The main idea of GBM is to focus and place the weights on incorrectly predicted results. While XGB is training, one tree trains the data set and assigns weights to incorrectly predicted records with errors, and the next tree of the same model learns the weighted data set and repeats the process of assigning weights. Moreover, GBM can quantify the contribution of features to the prediction results, such as feature importance. Particularly, XGB has the advantage of regularization and performance. It can perform parallel processing, regulate to avoid overfitting, is widely used for learning structured data, and has superior prediction performance.

Evaluation

We set the positive (1) label for discharge and the negative (0) label for hospitalization. To evaluate and compare the performance of candidate models, we used metrics including accuracy, sensitivity (recall for positive), specificity, precision, positive predictive value, negative predictive value, false-positive rate, and true-positive rate. When we monitored model training and validation, we used the F1-score to reflect imbalanced targets, the receiver operating characteristic (ROC) curve to find the optimal threshold, and the area under the ROC (AUROC) score to compare models.

To prevent overfitting the ML-based models and reduce biased results, we performed stratified, 5-fold cross-validation [46] illustrated in Figure 9. First, we randomly shuffled 63,261 PAIDs and divided them into 5 groups with approximately 12,000 people because we tried not to divide the records of a single patient into training (ie, plain box in Figure 9) and testing sets (ie, diagonal hatching box in Figure 9). Second, the first PAID group becomes the testing set, and the remaining groups become the training set in fold 1. We created fold 1 to fold 5 in a similar way to ensure equal division of the imbalanced targets (ie, the data set has true labels comprising 62.4% label 0 and 37.6%

label 1) across all folds. Besides, we split 25% of the training set as the validation set to tune the hyperparameters. Consequently, in each fold, we divided the data set into approximately 133,000 rows for the testing set and 535,000 rows for the training set (including the validation set). The ML-based models trained and tested all 5 folds.



Fold 1	Testing	Training	Training	Training	Training
Fold 2	Training	Testing	Training	Training	Training
Fold 3	Training	Training	Testing	Training	Training
Fold 4	Training	Training	Training	Testing	Training
Fold 5	Training	Training	Training	Training	Testing

Figure 9 Stratified 5-fold cross-validation to avoid overfitting

Individual Explainer for Outcome Assessment

Feature importance lists the features that the model considers prominent, and their contribution scores, in the process of training the data using the tree-based algorithm model. However, we considered XGB as the final model not only because of its high performance but also because of the access to the decision-making process inside the model. By approaching the trees, it is possible to describe the specific features and their influences that contribute to the prediction of each patient's daily prediction of discharge.

We demonstrate an individual explainer that can help in the interpretation of the XGB prediction results using a waterfall chart. Also called a bridge or cascade chart, it is a type of bar chart that portrays relative values and calculates the difference between adjacent values. It can show the positive or negative influence and gradual direction of the final discharge score.

To estimate values for individual explainers, we predicted the desired records with the trained XGB and obtained the contributions of all the features. The contribution refers to a feature's influence obtained by aggregating the scores that each feature contributes to all trees. Subsequently, we calculated the logistic value—*logistics (x) = 1 / (1 + e-x)* —of the feature's influence and the relative values required for the explainer. We selected the number of features to be displayed as 15, and the remaining 871 features were integrated and displayed simultaneously as others in the explainer.

Results

Data Characteristics

We created a data set that consisted of 669,667 records with 886 features, including diagnosis code, laboratory test results, physical information, medication, procedure, operation, PACS, and transfusion. Patients were admitted to cardiology or thoracic surgery, and their LOS ranged from 3 to 30 days. The average age of the patients was 61.03 (SD 13.42) years. The data set comprised 37.97% (254,254/669,667) women and 62.03% (415,413/669,667) men.

Performance of the ML-Based Predictive Models

Final ML-Based Model Selection

We experimented with the 5 ML-based models using 5 cross-validations. The AUROC score for each fold is listed in Table 9. The highest AUROC score for each fold is shown in italics, and the *support* column in Table 9 represents the number of each true label. Figure 10 shows the ROC curve plot; the area of the curve is represented by the AUROC and has a value between 0 and 1. The closer the AUROC score is to 1, the higher the model performance. XGB achieved the highest and a relatively stable score on all folds. Table 10 provides a comparison of the 5 ML-based models. All scores in Table 10 are the average values of the results and the SD in 5 folds, and the highest score for each metric is shown in italics. The specificity of logistic regression and support vector machine, which obtained 0.828, was the highest, but XGB achieved the highest in the rest of the metrics. Particularly, although the label of the data set was imbalanced, XGB scored 0.7 or higher for predicting label 1. Hence, we chose XGB as the final model to predict discharge probability.

Table 9 Evaluation by area under the receiver operating characteristic score of 5-fold cross-validation for each model.

	LR[a]	**SVM**[b]	**RF**[c]	**MLP**[d]	**XGB**[e]	**Support (0, 1)**
Fold 1	0.826	0.825	0.853	0.833	*0.866*[f]	(83,113, 50,188)
Fold 2	0.827	0.826	0.851	0.835	*0.868*	(83,538, 50,310)
Fold 3	0.824	0.824	0.850	0.821	*0.865*	(84,192, 50,585)
Fold 4	0.824	0.823	0.850	0.831	*0.864*	(83,969, 50,460)
Fold 5	0.822	0.821	0.848	0.834	*0.863*	(82,918, 50,394)
Mean	0.824 (0.002)	0.824 (0.002)	0.850 (0.002)	0.831 (0.005)	*0.865* (0.002)	N/A[g]

[a]LR: logistic regression.

[b]SVM: support vector machine.

[c]RF: random forest.

[d]MLP: multilayer perceptron.

[e]XGB: extreme gradient boosting.

[f]The italicized values refer to the highest score of each fold.
[g]N/A: not applicable.

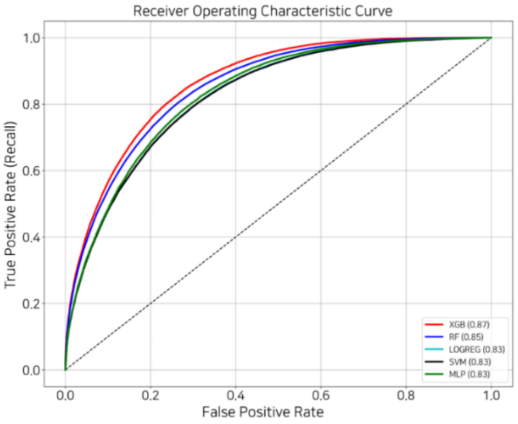


Figure 10 Receiver operating characteristic curve of the machine learning–based models. LOGREG: logistic regression; MLP: multilayer perceptron; RF: random forest; SVM: support vector machine; XGB: extreme gradient boosting.

Table 10 Comparison of the 5 machine learning–based models by metrics.

Model	**Values, mean (SD)**					
	ACC[a]	**Sen[b]**	**Spe[c]**	**PPV[d]**	**NPV[e]**	**AUROC[f]**
LR[g]	0.75 (0)	0.624 (0.005)	*0.828*[h] (0.004)	0.686 (0.005)	0.786 (0.005)	0.824 (0.002)
SVM[i]	0.75 (0)	0.624 (0.005)	*0.828* (0.004)	0.686 (0.005)	0.784 (0.005)	0.824 (0.002)
RF[j]	0.77 (0)	0.696 (0.005)	0.818 (0.004)	0.696 (0.005)	0.818 (0.004)	0.85 (0.002)
MLP[k]	0.758 (0.004)	0.642 (0.017)	0.822 (0.007)	0.686 (0.005)	0.792 (0.007)	0.831 (0.005)
XGB[l]	*0.782* (0.004)	*0.716* (0.005)	0.824 (0.005)	*0.71* (0)	*0.828* (0.004)	*0.865* (0.002)

[a]ACC: accuracy.
[b]Sen: sensitivity.
[c]Spe: specificity.
[d]PPV: positive predictive value.
[e]NPV: negative predictive value.

[f] AUROC: area under the receiver operating characteristic.
[g] LR: logistic regression.
[h] The italicized values refer to the highest score of each metric.
[i] SVM: support vector machine.
[j] RF: random forest.
[k] MLP: multilayer perceptron.
[l] XGB: extreme gradient boosting.

Figure 11 shows the relative feature importance of XGB sorted by gain score. The gain score refers to the average gain across all splits that the feature is used in. All the features used in the model have been replaced by their names used in the AMC. Except for the date-related feature, all other features that affected the model were found in all the tables. The features in the procedure table are substantially related to clinically critical situations. For example, the terms denoted with *(D)* are likely to mean a more severe state than others. The remaining features are also associated with CVDs or include primary examination and prescriptions during hospitalization.

However, because feature importance can only explain the model but not each patient, it is insufficient for use as an individual explainer for prediction. Depending on the patient’s condition, different features affect the daily probability of discharge. Therefore, we suggested an individual explainer that provides a patient-specific feature for daily prediction during hospitalization.

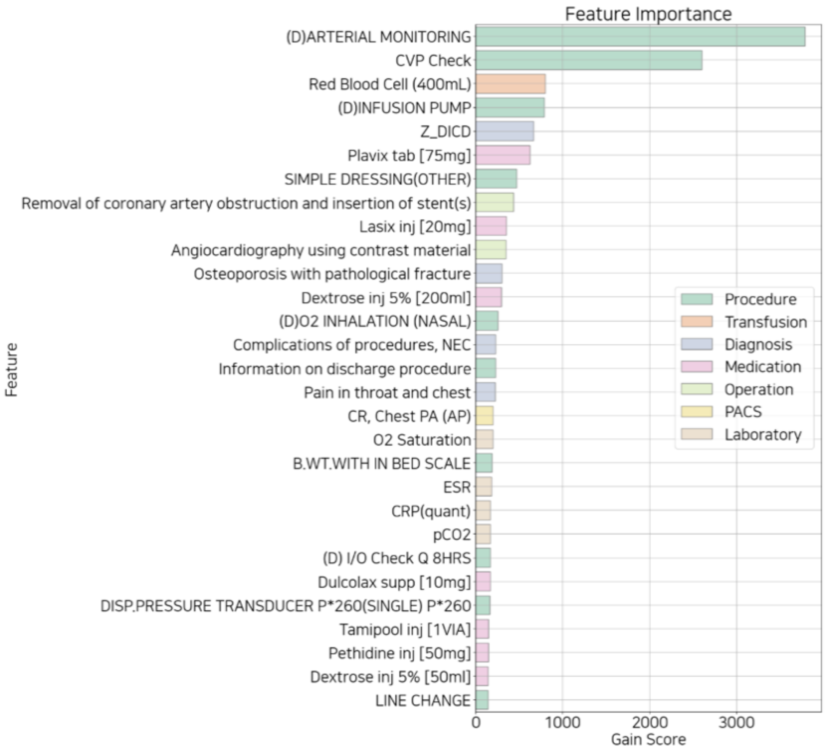


Figure 11 The feature importance sorted by gain score. B.WT.: body weight; CR: chest radiograph; CRP: C-reactive protein; CVP: central venous pressure; DISP: disposable;

ESR: erythrocyte sedimentation rate; I/O: intake and output; supp: suppository; inj: injection; NEC: necrotizing enterocolitis; PA: posteroanterior; PACS: picture archiving and communication system; Z_DICD: all diagnostic codes not selected for one-hot encoding.

Feature Reduction

Too many features tend to reflect negatively on the model performance; therefore, it was necessary to select an appropriate number of features. We performed recursive feature elimination with cross-validation (RFECV). This algorithm aims to identify the optimal number of features by comparing model performance while eliminating the features with low feature importance one at a time. RFECV returns the ranks and names of all features; we identified approximately 150 features with a rank of 1 by applying RFECV to our final model XGB. For performance comparison, we performed 5-fold cross-validation using the same data set with the same parameters. The number of features to be compared was 886 (all), 150 selected by RFECV, and the top 50 features in the model trained with the 150 selected by RFECV.

As shown in Figure 12 and Table 11, the performance difference between the model using all the features and the models with 150 and 50 features was only approximately 1% to 2.5% based on the AUROC score. This indicates that even with 83.1% to 94.4% of feature reduction, there is only a maximum performance difference of 2.5%. Therefore, a suitable number of features should be selected considering the situation in each hospital or the data characteristic.

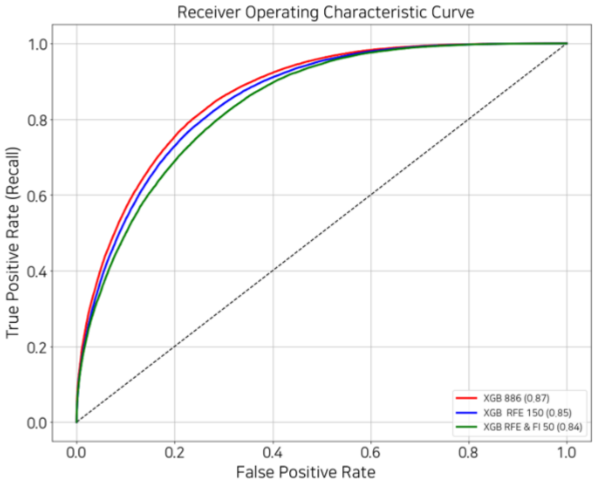


Figure 12 Receiver operating characteristic curve of the extreme gradient boosting models with the different number of features. FI: feature importance; RFE: recursive feature elimination; XGB: extreme gradient boosting.

Table 11 Evaluation by area under the receiver operating characteristic (AUROC) score of 5-fold cross-validation to select features.

Number of features	**Values, mean (SD)**					
	ACC[a]	**Sen[b]**	**Spe[c]**	**PPV[d]**	**NPV[e]**	**AUROC**
886 **(All)**	*0.782*[f] (0.004)	*0.716* (0.005)	*0.824* (0.005)	*0.71* (0)	*0.828* (0.004)	*0.865* (0.0018)
150 **(RFE[g])**	0.77 (0)	0.696 (0.005)	0.814 (0.005)	0.694 (0.005)	0.818 (0.004)	0.853 (0.0018)
50 **(RFE and FI[h])**	0.76 (0)	0.67 (0.006)	0.812 (0.004)	0.682 (0.004)	0.802 (0.004)	0.840 (0.00096)

[a] ACC: accuracy.
[b] Sen: sensitivity.
[c] Spe: specificity.
[d] PPV: positive predictive value.
[e] NPV: negative predictive value.
[f] The italicized values refer to the highest score of each metric.
[g] RFE: recursive feature elimination.
[h] FI: feature importance.

Explainer of Individual Prediction for Outcome Assessment

Overview

The predictive model classifies the data as *0* or *1* based on a threshold. The optimal threshold is the point where the sum of sensitivity and precision can be maximized simultaneously (in the ROC curve, true-positive rate and false-positive rate are proportional to each other). However, sensitivity and precision require trade-off against each other; therefore, decreasing FN increases sensitivity, and decreasing false positive increases precision. In other words, it is necessary to adjust for the appropriate threshold to suit the decision point of the hospital operation.

We presented the daily discharge score during hospitalization and the influence of the features by date through the explainer of individualized predictions. The following section includes a description and an example of our explainer for the sample data set, which represents one of the patients in the test set.

Discharge Score During Hospitalization

The sample data set consisted of the records of a patient with a PAID of 228,443 and an INNO of 2, hospitalized for 13 days and discharged on day 14. The patient's daily discharge score plot is depicted in Figure 13. The plot's x-axis represents the daily date excepted discharge date (ie, day 14) within the patient's hospitalization period, and the y-axis represents the probability of discharge (ie, discharge score). The model's optimal threshold was 0.39, indicated by a horizontal dotted line. The circle and the triangle represent the true labels *1* and *0*, respectively, and the size of the figure is proportional to the discharge score. The colors of the figure denote the results predicted by the model: red for positive prediction (label *1*, discharge) and blue for negative prediction (label *0*, admission).

For this sample, the model accurately predicted the discharge within 3 days. However, if the threshold is adjusted, the prediction results may change on dates 11 and 12. For example, if the current threshold rises slightly, *1* is applicable only for dates 12 and 13. This can be useful when trying to avoid false positive even if the false negative increases.

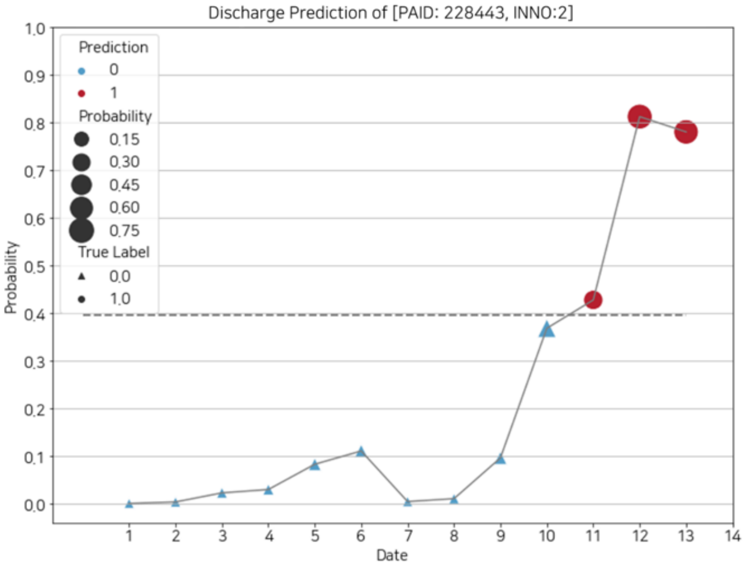


Figure 13 Daily discharge score of a patient's identification of 228,443 and patient's encounter number of 2. INNO: patient's encounter number; PAID: patient's identification.

Daily Feature Influence Score

Figures 14 and 15 describe the plot of feature influence for each day. The following is the basic description of the individual explainer: the x-axis of the plot is a score ranging from 0 to 1, and the y-axis represents the contributed features and the corresponding values that influenced the probability of discharge on that day. The threshold represented by the vertical dotted line is equal to the optimal threshold in Figure 14. The intercept, the plain blue box at the bottom of the y-axis, is a revised value reflecting that the number of each true label is imbalanced. The discharge probability, the gray box at the top of the y-axis, is the discharge score, which is the same as the probability in Figure 14. The width of each box corresponding to the feature refers to the absolute value of each score. The original score is indicated on the right side of the plot. The absolute value decreases from bottom to top, which means the contribution to the discharge score also decreases (the box of *others* is relatively wide because it is the sum of the scores of approximately 800 features, excluding the features below it). The red box with diagonal hatching represents each score of the feature that positively contributed to the discharge score and moves to the right. Conversely, the plain blue box represents negatively contributing feature scores and moves to the left.

To summarize, on the y-axis, from bottom to top, the features contributed to the prediction; the diagonal hatched red box to the right is positive, and the plain blue box to the left is negative.

Figure 14 shows the feature influence at day 7 with a low probability of discharge of 0.004, and Figure 15 shows day 12 with a high probability of 0.811. In Figure 14, *arterial monitoring=1* and *infusion pump=3* negatively affected the probability. In contrast, in Figure 15, *infusion pump=0* had a positive effect on probability. Because arterial monitoring and infusion pump are mainly prescribed for critical patients, both consist mostly of zeros in the data set. Therefore, displaying features and values together can help the medical staff interpret the plot intuitively. Moreover, each explainer may or may not have the features that appeared in the feature importance plot. This suggests that it is also necessary to identify features that contributed to individual patients rather than managing only the features of feature importance.

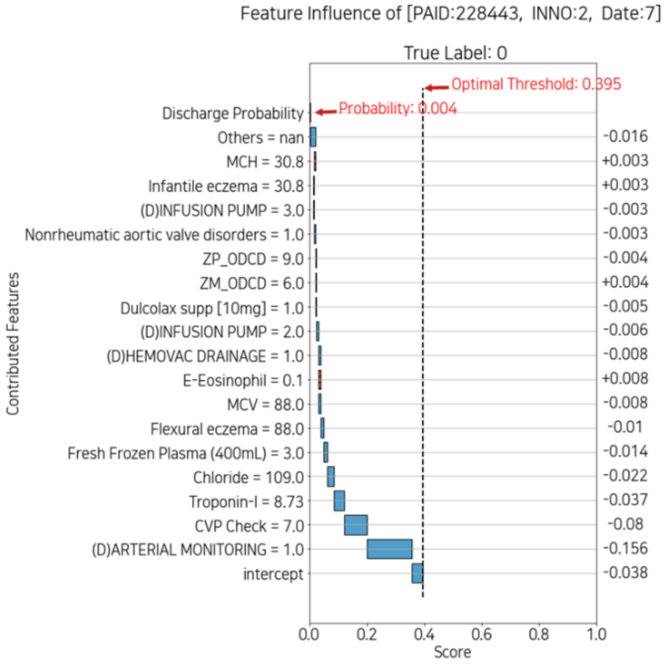


Figure 14 Feature influence with low probability of discharge date 7. CVP: central venous pressure; INNO: patient’s encounter number; MCH: mean corpuscular hemoglobin; MCV: mean corpuscular volume; PAID: patient’s identification; supp: suppository; ZM_ODCD: all medication codes not selected for one-hot encoding; ZP_ODCD: all procedure codes not selected for one-hot encoding.

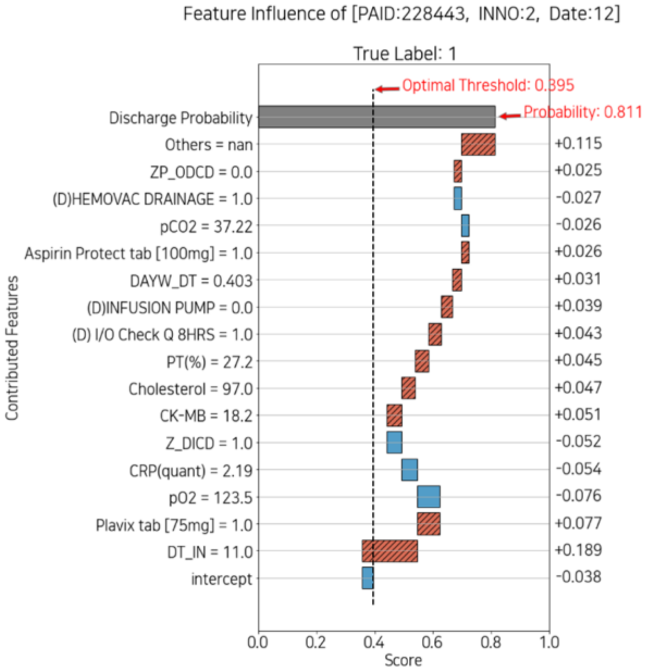


Figure 15 Feature influence with high probability of discharge on date 12. CK-MB: creatine kinase-myoglobin binding; CRP: C-reactive protein; DAYW_DT: integer feature of weekday; DT_IN: time since admission date in days; I/O: intake and output; INNO: the patient's encounter number; PAID: the patient's identification; PT: prothrombin time; Z_DICD: all diagnostic codes not selected for one-hot encoding; ZP_ODCD: all procedure codes not selected for one-hot encoding.

Outcome Assessment

Figure 5 shows the simulated impact in bed management applied with our predictive model and individual explainer. It is possible to recognize the probabilities of discharge of all patients for each ward every day. The paramount features and values that affect the discharge scores can be identified at once. It is informative for interpreting both high or low probability because the explainer implies the reasoning not only for discharge but also prolonged discharge. Similarly, it is possible to obtain information based on the expected discharge date of each patient, such as bed capacity in the near future. For the human and physical resources of the hospital to be used efficiently, future bed availability information can help reduce hospital costs through better management of beds and hospitalization reservations.

Discussion

Principal Findings

Investigations into bed management, which requires the use of hospital processes, and biomarker detection for patient-specific care, are actively pursued. In this study, we propose an ML-based predictive model to identify the discharge date for better bed management and the risk factors regarding discharge and CVDs. However, because each hospital has varying environmental variables, an algorithm that can consider them collectively was needed. Our study can contribute to improving the algorithm and supporting health care services. We have summarized the expectations of our predictive model and its explanation, along with its limitations.

First, we predicted the possibility of discharge to learn future information, but for the model to be practically applied, objective information about the current bed situation must be obtained. Currently, we are collecting bed information to combine it with the prediction results and optimize overall bed management. Consequently, our predictive model can be extended from ward-level up to hospital-level bed management. It may reduce the labor-intensive tasks for the medical team and the waiting time for patients.

Second, although our model provides adjustment of the optimal threshold according to the hospital circumstances, the ambiguity of decision-making because of results near the threshold exists, such as dates 10 and 11 in Figure 13. To solve this problem, there is a method that uses weighted average to make the result more conservative but reliable. Instead of using the probability returned by the model directly, it may be more useful to use it after weighting it for the past results, so that the target day reflects the weighted past results. It is just as necessary to produce reliable results as it is trying to explain the model and its internal features.

Finally, EHRs are longitudinal and sequential, but the sequence is different for each patient, and they do not have a regular interval. Consequently, we are preparing a preprocessing technique that can properly control the EHRs and reflect them in the model. Furthermore, compared with computer visualization, sequential data are relatively difficult to apply to XAI. Still, we are preparing explainable methods that are compatible with these data.

Conclusion

We proposed an individual explainer based on an ML-based predictive model, which provides the discharge probability and relative contributions of individual features. Our model can assist medical teams and patients in identifying individual and common risk factors in CVDs and support hospital administrators in improving the management of hospital beds and other resources.

Chapter 3. Deep Learning-based estimated Glomerular Filtration Rate Risk Prediction for Inpatients with Heart Failure

Introduction

Background

Cardiovascular diseases (CVDs) are one of the diseases with a steadily increasing number of patients. CVDs are largely divided into acute and chronic, and among them, acute CVDs can lead to death when disease occurs, so continuous and active prevention is important [47]. In addition, in the case of patients with chronic kidney diseases (CKDs), CVDs are often found, and it is shown that CKDs and CVDs are related [48,49].

For active prevention, many machine earning (ML)-based studies for managing CVDs are being conducted [50,51]. Using ML and deep learning (DL) to create and/or repurpose drug treatments for CVDs, perform diagnosis through CVD-related images and electrocardiogram analysis, etc. In addition, there are many disease-based studies that predict CVDs occurrence, recurrence, and re-hospitalization [52-55].

Among CVDs, heart failure (HF) is one of the acute diseases in which the heart fails to supply blood properly due to functional and structural disorders of the heart, and the recurrence rate is high. Coronary artery disease is known to be the main cause of HF. Indicators that have been found to be directly related to CVDs include creatine kinase-MB, Troponin I, and NT-Pro-BNP, but these are not often performed in laboratory tests [56].

Therefore, Using the estimated glomerular filtration rate (eGFR) calculated based on the creatinine level, which is one of the basic and widely measured tests, we developed a DL-based predictive model for the detection and monitoring of HF inpatient disease risk.

Objectives

The main contributions of this study can be summarized in the following steps:

First, we proposed a method to effectively convert the fragmentary stored electronic medical records (EMRs) into model training data. The popularization of EMRs provided convenience to both medical staff and patients and increased the efficiency of document management. [57] However, since EMRs was developed for document storage, it is difficult to grasp time-series information at once. Since most EMRs are stored heterogeneously, and even the data of the same patient is stored in clinical databases of different formats depending on the type of data, we proposed efficient pre-processing method to create learning dataset with time series for DL-based model.

Second, we developed a DL-based model to monitor heart failure inpatients by predicting the value and risk range of the eGFR. Since the glomerular filtration rate (GFR) is cumbersome to directly calculate as the amount of blood filtered by the kidneys in 1 minute, eGFR is calculated and used based on creatinine and demographics. [58,59] The unit of eGFR is ㎖/min/1.73 m2, and 90 or more indicates normal, 60 or more and less than 90 mild risks, 30 or more and less than 60 moderate risks, 15 or more and less than 30 severe risks, and less than 15 indicate profound risks. [60,61] The estimation methods include modification of diet in renal disease (MDRD) and chronic kidney disease epidemiology collaboration (CKD-EPI), and the CKD-EPI method is recommended when

GFR is 60 or higher. Several methods are being studied to improve the accuracy of the estimated calculated eGFR. If the eGFR is continuously reduced below the normal range, it could pose a risk to both the kidneys and the heart, so a model for predicting the level and range of eGFR was developed. Our model predicts the eGFR level after 12 hours, 24 hours, 36 hours, and 48 hours by using variables obtained during hospitalization and predicts the classification of the risk range of eGFR. It could assist the medical team in decision-making and support precision medicine.

Methods

We extracted data of hospitalized patients with HF, performed pre-processing, and created a dataset including time series to train a DL-based model to make predictions for eGFR. In addition, we selected features so that our model could be used in clinical practice and compared the performance. Finally, we visualized the outcomes of the model to explain results.

Description of patients

We extracted the data from CardioNet, a manually curated EMRs database specialized in CVDs. [39] There are 572,811 patients who had visited Asan Medical Center (AMC) with CVDs between January 1, 2000, and December 31, 2016. The AMC institutional review board approved the collection of CardioNet data and waived informed consent. CardioNet contains 27 tables on topics such as visitation, demographics, diagnosis, medication, and laboratory examination.

Most tables have common variables including patient identification (PAID), patient encounter number (INNO), the date of visitation or admission (INDT), and the date of discharge (OUDT). The KEY column, which concatenates the PAID and INNO columns, can connect the visitation table to other tables. Using the KEY column, we extracted the variables in each table to be analyzed.

The following is the list of extracted data:

- Visit table: PAID, INNO, KEY, date of visitation or admission, date of discharge, type of visit, medical department.
- Diagnosis table: International Classification of Diseases, Tenth Revision code of diagnosis.
- Laboratory test results table: date and code of pathology examination, and the results of the examination.
- Physical information table: patient’s age, height, weight, systolic and diastolic blood pressures, respiratory rate, pulse rate, BMI, body surface area, and date of measurements.
- Medication table: date and code of prescription.
- Echocardiography table: date, features, and results of echocardiography.

We extracted 12,195 hospitalization records of 8,418 anonymous patients who were hospitalized in the department of cardiology or thoracic surgery including hospitalization via the emergency department and who were diagnosed with HF. All patients’ lengths of stay (LOS) are between 3 days and 30 days because long-term patients are separately managed by the AMC.

Data Preprocessing

Baseline dataset

First, we created the baseline dataset using visits, emergency department visits, diagnosis, and laboratory test results tables. We included inpatients with a LOS of at least 3 days and less than 30 days to predict 12, 24, 36, and 48 hours later because hospitalizations of 31 days or longer were managed separately by the AMC.

Second, the start of each patient's hospitalization case in the baseline dataset was assigned as the earliest by comparing the hospitalization time in the visit table or the initial examination time in the laboratory test results table. Based on the emergency department visit table, for patients admitted via the emergency room, the emergency room admission time was designated as the start of the baseline dataset.

Third, after creating the starting index for each hospitalization case, data were generated one by one in a time step of 12 hours until the discharge date and all data in other tables were linked by the patient and each time step.

The following is a detailed description of the preprocessing method for each table:

Diagnosis table

The main diagnosed codes are cardiomyopathy and HF, and additional comorbidities include hypertension cerebral infarction, angina pectoris, acute myocardial infarction, chronic ischemic heart disease, pain in the throat and chest, and diabetes.

Laboratory test results table

Among the lab tests performed by heart failure patients during hospitalization, 62 tests performed by patients with more than 10 percent were selected as variables. (Excluding eGFR, which is the target of the DL-based model) Creatinine was 99.78%, potassium and sodium were 99.73%, and total CO2 was 99.7%. After dividing the laboratory test results into 12-hour increments, the values were filled in with the corresponding time step for each patient and hospitalization. If there are multiple values in one time step, the median value is substituted. If there is a value even for one time step for each variable, back-fill, and forward-fill are performed once. In the case where the laboratory test was never performed, all were replaced with -1.

Physical information table

We included height, weight, systolic blood pressure, diastolic blood pressure, respiration rate, and heart rate from the physical information table. In order to remove error data, only 0.001 to 0.999 percent of the data were used. As in the case of the laboratory results test table, back-fill and forward-fill were performed when there was a result at least once in the same hospitalization case, and -1 was replaced when there was no result at all.

Medication table

Present we included 86 drugs prescribed by inpatients more than 1% among the list of cardiac-related medications classified by cardiologists, Additionally, 79 medications prescribed by more than 10% of CardioNet patients were combined and duplicates were removed to select a total of 149 medication codes.

Past For each patient, drugs prescribed for 28 days or longer that are expected to be taken for a long time within the past 5 years were selected. The medication history was filled in binary into 30 categories divided by the clinician. These values are all assigned the same value to the same hospitalization case, unlike the currently prescribed medication.

Echocardiography

The echocardiography from CardioNet was created by structuring the reading statement of echocardiography. Among the results of hospitalized patients, all variables that did not have a value greater than 50% were removed, and the remaining variables were grouped so that they had the same value in one hospitalization and filled with the same data.

Target criteria

The target of the DL-based model is eGFR, and the estimation methods used in AMC are MDRD and CKD-EPI. Both were calculated based on creatinine, and the median value of MDRD and CKD-EPI were selected as regression targets. If there is no value divided according to the time step, the value is substituted to have a linear value using interpolation. As a classification target, five multi-class labels were created according to severity. The target ratio of the dataset is shown in Table 12.

Table 12 Labeling based on severity of eGFR. eGFR: estimated glomerular filtration rate

Label	Severity	Range of eGFR	Number of labels
0	Profound	eGFR < 15	26,781
1	Severe	15 ≤ eGFR < 30	25,059
2	Moderate	30 ≤ eGFR < 60	61,870
3	Mild	60 ≤ eGFR < 90	71,878
4	Normal	90 ≤ eGFR	58,324
Total	-	-	243,912

As a result, we extracted 316 features from seven tables from CardioNet and created a dataset of 243,912 rows from 12,195 hospitalization cases of 8,418 patients with HF.

DL-Based Predictive model

Learning dataset

Before transforming the learning data for the DL-based model, all hospitalization cases (i.e., KEY) were shuffled. When mixing after creating learning data, the time order of each hospitalization is mixed, so the hospitalization case was shuffled in advance. In addition, since our model predicts 1, 2, 3, 4 steps (i.e., 12, 24, 36, 48 hours) after looking at identical 2-time steps, we created different learning datasets for each time step.

Learning datasets can be divided into model training (including validation) and testing. We divided the learning datasets for model training (including validation) and evaluation at 0.8 and 0.2 ratios, and within model training, training and validation ratios were divided into 0.75 and 0.25. Detailed values are shown in Table 13.

Table 13 The number of rows in training, validation and testing per time step

Time step	Learning datasets	Model training	Model testing
0 (12 hours)	219,522	176,148	43,374
1 (24 hours)	207,366	166,424	40,942
2 (36 hours)	195,434	156,876	38,558
3 (48 hours)	183,653	147,454	36,199

DL-Based models

DL is a type of machine learning that uses an artificial neural network (ANN) with successive layers to gradually learn characteristics or representations of data within layers [62]. To consider the time series, we adopted the Long Short-Term Memory (LSTM) of recurrent neural networks (RNN) [63], which is one of the DL-based models. The predictive model we developed learns 24 hours of data, predicts eGFR levels after 12, 24, 36, and 48 hours and predicts one of five risk labels.

Evaluation

we evaluated the regression prediction results using mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R-squared). The classification prediction results were evaluated based on accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUROC). However, there is a limitation in that one label can be compared with the rest in the case of multi-class.

Results

Characteristics

As a result, we created learning datasets that consisted of 243,912 rows with 316 features, including diagnosis code, laboratory test results, physical information, present and past medication, and echocardiography. The data set consisted of 12,195 hospitalized cases (i.e., 3,493 hospital admissions (28.6%) and 8,702 hospitalizations via emergency room (71.4%)) of 8,418 patients (i.e., 3,723 women (44.2%) and 4,695 men (55.8%)). The average age of the patients was 63.45(SD 16.29) years. In addition, about 8,702 cases of hospitalization via the emergency room and 3,493 cases of just hospitalization.

Performance of the DL-Based model

Figure 16 is the structure of our adopted DL-based model. We stacked the LSTM layer in three layers and performed prediction by dividing it into classification and regression. Additionally, we defined score – *1-accuracy + MAE* – to find the suitable weighted combination of regression and classification, and the lowest score was found to be 0.8:0.2. The scores for all combinations are shown in Table 14.

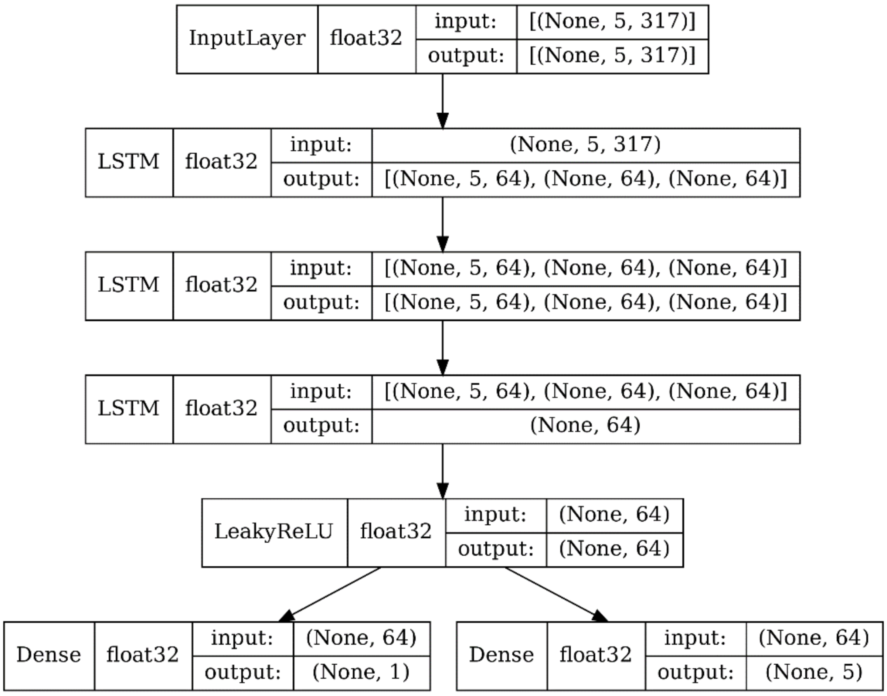


Figure 16 The structure of DL-based predictive model for eGFR. DL: deep learning, eGFR: estimated glomerular filtration rate

Table 14 Comparison of MSE, MAE, and Accuracy by combination of weighted loss combination. MSE: mean squared error, MAE: mean absolute error

The ratio of weighted loss		**MSE**	**MAE**	**Accuracy**	**Score**
Regression	**Classification**				
0	1.0	6042.819	63.967	0.784	64.184
0.1	0.9	165.759	5.807	0.854	5.953
0.2	0.8	165.811	5.983	0.851	6.132
0.3	0.7	160.202	5.851	0.854	5.997
0.4	0.6	160.105	5.639	0.855	5.784
0.5	0.5	160.35	5.754	0.858	5.896
0.6	0.4	159.991	5.896	0.856	6.04
0.7	0.3	172.835	6	0.851	6.149
0.8	0.2	**158.134**	**5.571**	**0.864**	**5.707**
0.9	0.1	163.654	5.791	0.86	5.931
1.0	0	166.13	5.772	0.221	6.551

Evaluation of DL-based predictive models

The results of the DL-based predictive models are listed in Table 15. The model that looked at 2-time steps and predicted values after 12 hours showed the highest R-squared and accuracy, and the lowest MSE and MAE. This shows that the shorter the time step (i.e., skipping time), the higher the prediction performance and indicate that there is a trade-off between the more distant future and the performance.

Table 15 Comparison of MSE, MAE, R-squared and Accuracy of models by time step MSE: mean squared error, MAE: mean absolute error, R-squared: coefficient of determination

Time step	**MSE**	**MAE**	**R-Squared**	**Accuracy (All)**
0 (12 hours)	169.626	5.82	0.912	0.851
1 (24 hours)	263.051	8.555	0.865	0.783

2 (36 hours)	380.315	10.794	0.808	0.717
3 (48 hours)	457.098	12.694	0.773	0.676

Table 16 is a classification report for risk classification prediction, including precision, recall, and F1-score for one class and all of the rest classes. (i.e., class 0 versus class 1, class 2, class 3, and class 4) Although class 0 was a label for profound HF, the score was higher than that of other classes, followed by class 4. In addition, Figure 17 depicts the ROC of predictive classification for each time step model. The performance of class 0 is the highest compared to other classes, and the performance of class 3, labeled as mild risk, is the lowest.

Table 16 Classification report of models by time step

Timestep **0** **(12 hours)**	**Class**	**Precision**	**Recall**	**F1-score**	**Support**
	0	0.92	0.97	0.94	5123
	1	0.86	0.77	0.81	4519
	2	0.86	0.84	0.85	10922
	3	0.79	0.83	0.81	12507
	4	0.88	0.87	0.87	10303
	Accuracy			0.85	43374
Timestep **1** **(24 hours)**	**Class**	**Precision**	**Recall**	**F1-score**	**Support**
	0	0.92	0.92	0.92	4825
	1	0.79	0.66	0.72	4262
	2	0.78	0.76	0.77	10254
	3	0.7	0.78	0.73	11749
	4	0.83	0.8	0.82	9852
	Accuracy			0.78	40942
Timestep **2** **(36 hours)**	**Class**	**Precision**	**Recall**	**F1-score**	**Support**
	0	0.92	0.84	0.88	4539
	1	0.66	0.6	0.63	4003

	2	0.72	0.66	0.69	9609
	3	0.62	0.72	0.67	11026
	4	0.78	0.76	0.77	9381
	Accuracy			0.72	38558
Timestep 3 (48 hours)	**Class**	**Precision**	**Recall**	**F1-score**	**Support**
	0	0.9	0.83	0.86	4257
	1	0.62	0.55	0.58	3754
	2	0.67	0.64	0.65	8980
	3	0.58	0.65	0.61	10310
	4	0.73	0.72	0.72	8898
	Accuracy			0.68	36199

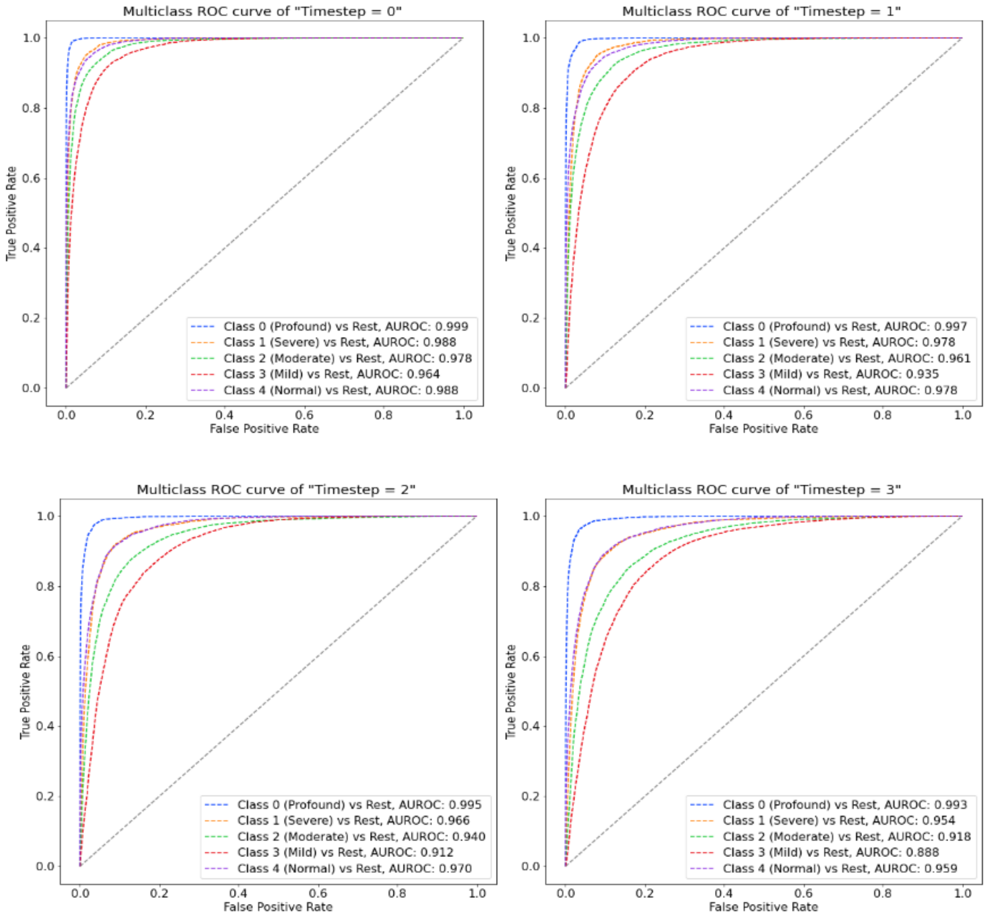


Figure 17 The ROC for each (time step) model including 5 classes. The score next to each class means AUROC score of one class versus the rest. ROC: receiver operating characteristic, AUROC: the area under ROC

Visualization of outcomes

The overall eGFR graph shown in Figure 18, represented the results of predictions of all time step models for one hospitalization data of a sample patient who was hospitalized for 6-time steps (i.e., about 3 days). The black circle and grey line mean the original data of eGFR. Each red triangle, green square, blue pentagon, and pink star indicate 12 hours, 24 hours, 36 hours, and 48 hours after predictions. The background presents the risk according to the eGFR risk classification stage. The red triangles are results of predicting immediately after 12 hours based on 24-hour learning data, which are following a trend closer to the original value than other results.

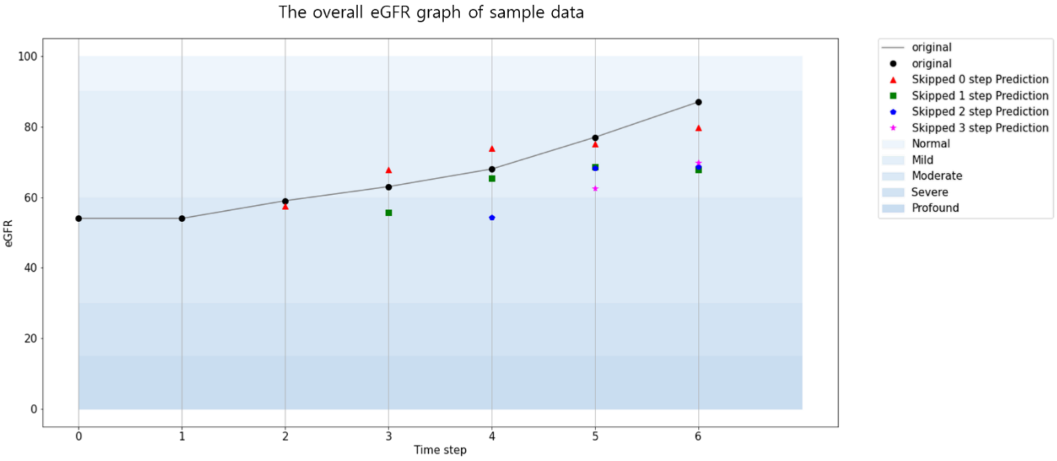


Figure 18 The overall eGFR graph of sample data. eGFR: estimated glomerular filtration rate

Figure 19 was implemented as a model to predict after 12 hours (a) and 24 hours (b) for a sample hospitalization as in Figure 18. The x-axis and y-axis indicate the time step and value of eGFR, and the colors of the background present the risk range of eGFR as in Figure 18. The black star and grey lines represent the two-time original value of eGFR as input to the DL-based model. The blue stars present the target data that is the correct answer, while the red circles are the prediction results of models. It is shown that the original value and the predicted value are slightly different in (b) compared to (a).

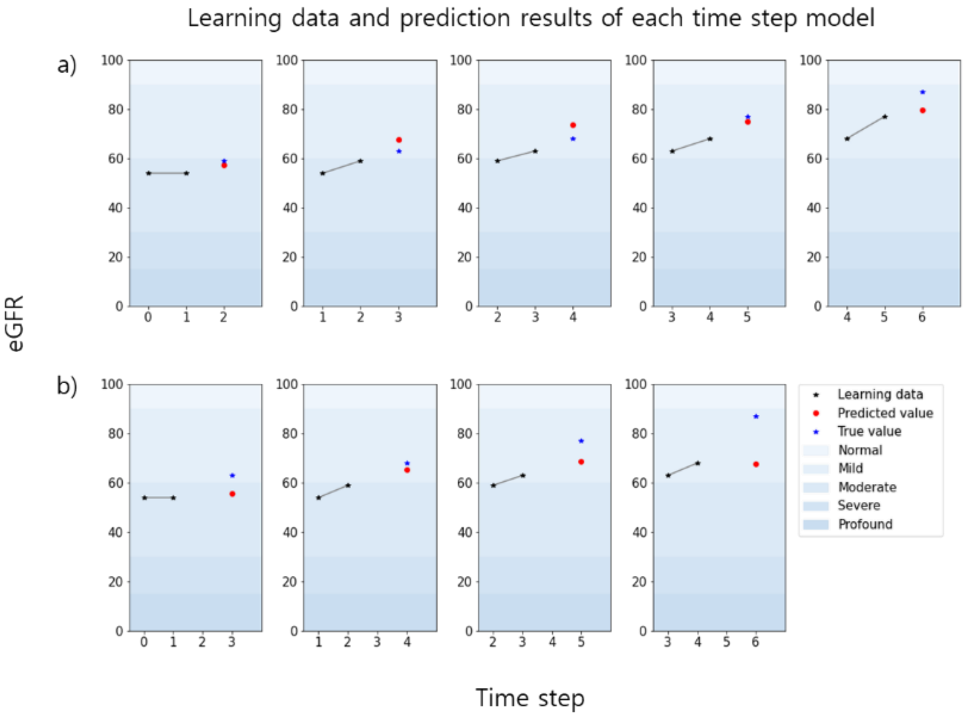


Figure 19 The learning data and prediction results of each time step model for sample data (a) 12 hours later prediction model, b) 24 hours later prediction model)

As a result, we extracted and pre-processed the heterogeneous data from EMRs, and developed a model to predict the level and risk range of eGFR based on LSTM, one of the time series predictive models, and achieved MSE of 169.626, MAE of 5.82, and R-squared of 0.912 for regression performance through DL-based models. For the risk range classification problem of eGFR, models accomplished an accuracy of 85.1%. In addition, we presented the overall outcome for each patient's eGFR within the hospital stay and divided graphs for each time step. Due to the visualization process, it is expected that it could contribute to identifying and utilizing data in clinical practice.

Discussion

First, there are several methods for calculating eGFR based on creatinine, and many studies on more accurate estimation methods are in progress. Since this study utilized the median values of MDRD and CKD-EPI, we thought that a more precise estimation calculation method would be helpful for accurate prediction.

Second, because we required the data measured twice every 12 hours, the study was conducted on only hospitalized patients to obtain data of at least 6-time steps. It is necessary to expand the study to model development that shows elaborate prediction with fewer 6-time laboratory tests, and it is expected to apply not only to inpatients but also to outpatients.

Third, it is difficult to check the inside of the model due to the basis of the DL-based model. There are many ML-based algorithm studies to improve performance. However, research to study the inside of the model by understanding the operating principle of the model is also required. Therefore, it is mandatory to extend the study which can explain the model by finding the importance of each time series or feature by utilizing other algorithms or combinations with other models.

We experimented with ML and DL-based models together to find an appropriate number of features that can be applied in clinical practice. Although the performance improved and the utility increased by reducing the number of features, there are still difficulties in looking at the inside of the DL-based models.

Conclusion

We conducted the effective pre-processing method for generating sequential data from EMRs. We developed the DL-based predictive model providing the value and risk of eGFR for inpatients with HF. Additionally, we presented overall and divided graph by time step which could support the medical team and patients in managing the risk of HF and CVDs in advance.

Conclusions

In the first study, we built a large-scale and integrated CardioNet database to apply AI technology in CVDs for the detection of risk factors, development of predictive models for early diagnosis, and improving the care of patients. First, we obtained the EHR data with approval from the IRB of AMC and UUH. Second, we processed structured and unstructured data appropriately using medical expertise to generate data that can be directly applied to the AI model. Finally, we standardized and validated the data in CardioNet to allow multi-centered research. CardioNet can contribute to the early prediction of cardiac problems and promote further CVD-related research.

In the second study, we have proposed an ML-based model to predict the daily discharge probability for each patient and demonstrated the individual explainer for any date during hospitalization, along with the reasonable contributing features. Our XGB model accomplished an AUROC of 0.865 and represented the simulated bed management based on explainable features. It could assist the medical team and patients in identifying the individual and common risk factors in CVDs and support hospital administrators in improving the management of hospital beds and other resources.

In the third study, we proposed an effective pre-processing method for generating sequential data from EMRs. In addition, we developed the DL-based model to predict the value and risk of eGFR for inpatients with HF. Our sequential DL-based model accomplished an MSE of 169.626, MAE of 5.82, an R-squared of 0.912 for regression, and an accuracy of 85.1% for classification. Additionally, we presented the overall graph and each graph by timestep including information on eGFR. The results of our DL-based predictive model could support the medical team and patients in managing the risk of HF and CVDs in advance.

Finally, we built a database specialized for CVDs, pre-processed EMRs, and developed the ML and DL-based predictive models for inpatients with CVDs. In addition, we visualized the outcomes of models to try to provide informative and explainable data to medical teams and patients. These three studies could support improving the quality of healthcare and hospital management for realizing digital healthcare.

References

1. Evans RS, Lloyd JF, Pierce LA. Clinical use of an enterprise data warehouse. In: AMIA Annual Symposium Proceedings; 2012. p. 189
2. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet. 2019;394(10201):861–867.
3. Organization WH. Cardiovascular diseases (CVDs) fact sheet. World Health Organization. 2017;
4. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature medicine. 2019;25(6):954–961.
5. Jun TJ, Kweon J, Kim YH, Kim D. T-Net: Nested encoder-decoder architecture for the main vessel segmentation in coronary angiography. Neural Networks. 2020;
6. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89–94.
7. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering. 2018;2(3):158.
8. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging. 2014;34(10):1993–2024.
9. Hollon TC, Pandian B, Adapa AR, Urias E, Save AV, Khalsa SSS, et al. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. Nature medicine. 2020;26(1):52–58.
10. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019;572(7767):116–119.
11. Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In: Advances in Neural Information Processing Systems; 2016. p. 3504– 3512.
12. Harris ZS. Distributional structure. Word. 1954;10(2-3):146–162.

13. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics. 2015; 216:574.

14. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics. 2006; 121:279.

15. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clinical chemistry. 2003;49(4):624–633.

16. Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT professional. 2005;7(5):17–23.

17. Seo MH, Lee WY, Kim SS, Kang JH, Kang JH, Kim KK, et al. 2018 Korean Society for the Study of Obesity guideline for the management of obesity in Korea. Journal of obesity &amp; metabolic syndrome. 2019;28(1):40.

18. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–3119.

19. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014. p. 1532–1543.

20. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. arXiv preprint arXiv:180205365. 2018;

21. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. Jama. 2011;306(8):848– 855.

22. Wei Y, Yu H, Geng J, Wu B, Guo Z, He L, Chen Y. Hospital efficiency and utilization of high-technology medical equipment: A panel data analysis. HPT, 2018; 7(1): 65-72. DOI: 10.1016/j.hlpt.2018.01.001

23. Novati R, Papalia R, Peano L, Gorraz A, Artuso L, Canta MG, et al. Effectiveness of an hospital bed management model: results of four years of follow-up. Ann Ig, 2017; 29(3): 189-196. DOI: 10.7416/ai.2017.2146

24. Barnes S, Hamrock E, Toerper M, Siddiqui S, Levin S. Real-time prediction of inpatient length of stay for discharge prioritization. JAMIA, 2016; 23(e1): e2-e10. DOI: 10.1093/jamia/ocv106

25. World Health Organization. Cardiovascular diseases (CVDs). 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [accessed June 11, 2021]

26. Bachouch RB, Guinet A, Hajri-Gabouj S. An integer linear model for hospital bed planning. Int. J. Prod. Econ, 2012; 140(2): 833-843. DOI: 10.1016/j.ijpe.2012.07.023

27. Troy PM, Rosenberg L. Using simulation to determine the need for ICU beds for surgery patients. Surgery, 2009; 146(4): 608-620. DOI: 10.1016/j.surg.2009.05.021

28. Turgeman L, May JH, Sciulli R. Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission. ESWA, 2017; 78: 376-385. DOI: 10.1016/j.eswa.2017.02.023

29. Shimabukuro DW, Barton CW, Feldman MD, et al. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respiratory Research, 2017; 4: e000234. DOI: 10.1136/bmjresp-2017-000234

30. Fei M, Limin Y, Lishan Y, David DY, Weifen Z, Length-of-Stay Prediction for Pediatric Patients With Respiratory Diseases Using Decision Tree Methods, IEEE Journal of Biomedical and Health Informatics, 2020; 24(9): 2651-2662. DOI: 10.1109/JBHI.2020.2973285

31. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digital Med, 2018; 1(1): 1-10. DOI: 10.1038/s41746-018-0029-1

32. Tamarappoo BK, Lin A, Commandeur F, McElhinney PA, Cadet S, Goeller M, et al. Machine learning integration of circulating and imaging biomarkers for explainable patient-specific prediction of cardiac events: A prospective study. Atherosclerosis, 2021; 318: 76-82. DOI: 10.1016/j.atherosclerosis.2020.11.008

33. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 2018; 6: 52138-52160. DOI: 10.1109/ACCESS.2018.2870052

34. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen M J, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun, 2020; 11(1): 1-11. DOI: 10.1038/s41467-020-17431-x

35. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision, 2017; pp. 618-626

36. Jun TJ, Eom Y, Kim D, Kim C, Park JH, Nguyen HM, et al. TRk-CNN: Transferable Ranking-CNN for image classification of glaucoma, glaucoma suspect, and normal eyes. ESWA, 2021; 115211. DOI: 10.1016/j.eswa.2021.115211

37. Onishi K. Total management of chronic obstructive pulmonary disease (COPD) as an independent risk factor for cardiovascular disease. Journal of cardiology, 2017; 70(2): 128-134. DOI: 10.1016/j.jjcc.2017.03.001

38. Levin S, Barnes S, Toerper M, et al. Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay. BMJ Innovations, 2021; 7:414-421. DOI: 10.1136/bmjinnov-2020-000420

39. Ahn I, Na W, Kwon O, Yang DH, Park GM, Gwon H, et al. CardioNet: a manually curated database for artificial intelligence-based research on cardiovascular diseases. BMC Med Inform Decis Mak, 2021; 21(1): 1-15. DOI: 10.1186/s12911-021-01392-2

40. 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. DOI: 10.1016/S1532-0464(03)00034-0

41. Cortes C, Vapnik V. Support-vector networks. Mach Learn, 1995; 20(3): 273-297. DOI: 10.1007/BF00994018

42. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn, 2002; 46(1): 389-422. DOI: 10.1023/A:1012487302797

43. Breiman L. Random forests. Mach Learn, 2001; 45(1): 5-32. DOI: 10.1023/A:1010933404324

44. Yan H, Jiang Y, Zheng J, Peng C, Li Q. A multilayer perceptron-based medical decision support system for heart disease diagnosis. ESWA, 2006; 30(2): 272-281. DOI: 10.1016/j.eswa.2005.07.022

45. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. DOI: 10.1145/2939672.2939785

46. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai, 1995; 14(2): 1137-1145.

47. Mosterd A & Hoes AW. Clinical epidemiology of heart failure. *heart*, 2007; *93*(9), 1137-1146.

48. Elsayed EF, Tighiouart H, Griffith J, Kurth T, Levey AS, Salem D, ... & Weiner DE. Cardiovascular disease and subsequent kidney disease. *Archives of internal medicine*, 2007; *167*(11), 1130-1136.

49. Wright J & Hutchison A. Cardiovascular disease in patients with chronic kidney disease. *Vascular health and risk management*, 2009; *5*, 713.

50. Deo RC. Machine learning in medicine. *Circulation*, 2005; *132*(20), 1920-1930.

51. Johnson KW, Torres SJ, Glicksberg BS, Shameer K, Miotto R, Ali M, ... & Dudley JT. Artificial intelligence in cardiology. *Journal of the American College of Cardiology*, 2018; *71*(23), 2668-2679.

52. Mathur P, Srivastava S, Xu X, & Mehta JL. Artificial intelligence, machine learning, and cardiovascular disease. *Clinical Medicine Insights: Cardiology*, 2020; *14*, 1179546820927404.

53. Dinh A, Miertschin S, Young A, & Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC medical informatics and decision making*, 2019; *19*(1), 1-15.

54. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, ... & Jethwani K. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC medical informatics and decision making, 2018; 18(1), 1-17.

55. Choi E, Schuetz A, Stewart WF, & Sun J. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 2017; *24*(2), 361-370.

56. Landesberg G, Shatz V, Akopnik I, Wolf YG, Mayer M, Berlatzky Y, ... & Mosseri M. Association of cardiac troponin, CK-MB, and postoperative myocardial ischemia with long-term survival after major vascular surgery. *Journal of the American College of Cardiology*, 2003; *42*(9), 1547-1554.

57. Evans RS. Electronic health records: then, now, and in the future. *Yearbook of medical informatics*, 2016; *25*(S 01), S48-S61

58. Stevens LA, Coresh J, Greene T, & Levey AS. Assessing kidney function—measured and estimated glomerular filtration rate. *New England Journal of Medicine*, 2006; *354*(23), 2473-2483.

59. Inker LA, Schmid CH, Tighiouart H, Eckfeldt JH, Feldman HI, Greene T, ... & Levey AS. Estimating glomerular filtration rate from serum creatinine and cystatin C. *New England Journal of Medicine*, 2012; *367*(1), 20-29.

60. National Kidney Foundation. Estimated Glomerular Filtration Rate (eGFR). 2022. https://www.kidney.org/atoz/content/gfr [accessed April 19, 2022]

61. 대한신장학회. 신장 관련 검사. 2022. https://www.ksn.or.kr/general/about/check.php [accessed April 19, 2022]

62. LeCun Y, Bengio Y, & Hinton G. Deep learning. *nature*, 2015; *521*(7553), 436-444

63. Hochreiter S, & Schmidhuber J. Long short-term memory. Neural computation, 1997; 9(8), 1735-1780.

국문 요약

최근 의료기관에 저장되는 임상 데이터는 헬스케어 관련 기술의 발전으로 급증하고 있다. 전자의무기록은 임상 데이터 중의 하나로 환자의 다양한 진료 기록을 포함하고 있다. 이러한 환자의 기록은 개인정보보호로 인해 활용되지 못하였지만, 가명화나 익명화 등의 비식별화 과정을 통해 후향적 연구를 진행할 수 있게 되었다. 전자의무기록을 활용한 후향적 연구는 여러 위험 예측 연구를 수행할 수 있고, 실사용증거 (Real-World Evidence) 로 활용될 수 있다.

심혈관질환은 여러 동반질환을 수반하는 급성 및 만성 질환 중 하나로, 지속적이고 적극적인 관리가 필요하며, 이를 위해 인공지능 기반 연구들이 많이 수행되고 있다. 이러한 환자 관리를 지원하기 위해 다음 세가지 연구를 계획하였다. 첫째, 향후 심혈관질환 연구에 지속적으로 활용될 수 있는 심혈관질환 특화 데이터 베이스를 구축하는 데 목적이 있다. 둘째, 구축한 데이터 베이스를 활용하여 심혈관질환 관련 예측 연구들을 수행하기 위해 머신러닝 기반 모델을 개발하였다. 이 모델은 입원 환자의 퇴원 예측을 수행하여 효율적인 병원 자원 활용 지원에 그 목적이 있다. 셋째, 동일한 데이터 베이스를 활용하여 추정 사구체 여과율을 예측하는 딥러닝 기반 모델을 개발하여, 입원 환자의 심부전 위험을 감지하는 데 목적이 있다.

첫째, 적극적인 관리가 필요한 심혈관질환과 관련된 임상 데이터는 기본적인 외래, 입원 데이터를 비롯하여 심장초음파나 운동부하검사와 같은 다양한 특수 검사를 포함한다. 이러한 다양한 정형 데이터와 비정형 데이터를 통합하여 심혈관질환 관련 의료 정보학 연구 수행에 도움이 되고자 하여, 심혈관 질환 특화 데이터 베이스를 구축하였다. 익명화 된 데이터를 추출하고 임상적으로 수용가능한 기준에 따라 이상치 및 오류 데이터를 제거하였으며, 자연어처리 기법을 통해 문장 형태의 판독 결과지 등의 비정형 데이터를 구조화하여 심혈관질환 특화 데이터 베이스를 구축하였다. 구축된 데이터베이스는 전자의무기록 분석의 유용성을 높일 수 있으며, 2 차적 파생 연구를 지원할 수 있다.

둘째, 기 구축된 심혈관질환 특화 데이터베이스에서 심혈관질환 관련 입원 환자 데이터를 추출하여 머신러닝 기반 예측 연구를 수행하였다. 본 연구는 심혈관질환으로 입원한 환자들의 퇴원 가능성을 예측하였으며, 개인화된 설명자를 통해 각 환자의 입원 건 별 위험 요인 및 퇴원 가능성을 시각화 하였으며, 시뮬레이션된 자료를 통해 퇴원 가능성을 예측한 연구가 병원 프로세스 개선에 도움이 될 수 있음을 제시하였다.

셋째, 기 구축된 심혈관질환 특화 데이터베이스에서 심혈관질환 중 심부전으로 입원한 환자 데이터를 추출하고, 추정 사구체 여과율을 예측하여 질병 위험을 제시하는 딥러닝 기반 모델 개발 연구를 수행하였다. 본 연구에서는 전자의무기록을 딥러닝 기반 모델이 학습할 수 있는 시계열 학습 데이터로 변환하였으며, 추정 사구체 여과율을 예측하여, 심부전 및 만성신장질환을 앓고 있는 환자의 추정 사구체 여과율의 하락 등의 위험을 알려주는 예측 모델을 개발하였다. 추가적으로 각 기준 시간대에 따른 추정 사구체 여과율의 직관적으로 파악할 수 있는 시각화자료를 제시하였다.

결과적으로, 전자의무기록을 활용한 후향적 연구를 효율적으로 진행하기 위해 심혈관질환 특화 데이터베이스를 구축하였으며, 머신러닝 및 딥러닝 기반 모델 개발을 통해 다양한 예측 연구를 수행하였다. 본 연구를 확장하여 보다 정교한 모델 개발을 수행한다면, 의료의 질 향상 및 개인 맞춤형 디지털 헬스케어의 실현에 기여할 수 있을 것이라 생각된다.

첫번째 연구는 "CardioNet: a manually curated database for artificial intelligence-based research on cardiovascular diseases"의 제목으로 2021 년 1 월 28 일에 BMC Medical Informatics and Decision Making 에 출판되었으며, 두번째 연구는 "Machine Learning‑Based Hospital Discharge Prediction for Patients With Cardiovascular Diseases: Development and Usability Study"의 제목으로 2021 년 11 월 17 일에 JMIR medical informatics 에 출판되었다.