공학석사 학위논문

Disease Prediction Model Development Using Unstructured data from EMR and Artificial Intelligence

비정형 전자의무기록과 인공지능을 활용한 질환 예측 모델 개발

울산대학교 대학원
의 과 학 과
한 지 예

Disease Prediction Model Development

Using Unstructured data from EMR

and Artificial Intelligence

지도교수 김영학

이 논문을 공학석사학위 논문으로 제출함

2024년 02월

울산대학교 대학원

의과학과

한지예

한지예의 공학석사학위 논문을 인준함

심사위원 이 계 화 (인)

심사위원 김 영 학 (인) 김영학인

심사위원 전 태 준 (인) 전태준인

울 산 대 학 교 대 학 원

2024 년 02 월

Abstract

Background

Clinical research utilizing electronic medical records encompasses diverse forms of medical data. Moreover, owing to recent advancements in natural language processing technology, there is a burgeoning interest in investigating text data embedded within electronic medical records. This research contributes to real-world evidence (RWE) in authentic clinical settings and, when coupled with artificial intelligence technology, has the potential to make significant contributions to various domains, including disease prediction and medical decision support.

Objectives

First, we aim to demonstrate the benefit of early treatment by studying the association between achievement of early LDL-C goal and recurrence of MACE and healthcare resource utilization (HRU) in high-risk ASCVD patients through electronic health records. Second, we aim to leverage unstructured text data to develop a prediction model for disease progression in ICUS patients, identify associated risk factors, and provide clinical insights into patient management.

Methods

First, patients with cardiovascular disease were defined based on clinical evidence and then divided into two groups depending on whether they achieved the early LDL-C reduction goal. The results of the analysis regarding the risk ratio of recurrent major cardiovascular events (MACE) and the frequency of medical resource utilization were supported through statistical validation. Second, we utilized unstructured EMR text data to construct a dataset that captures the characteristics of Idiopathic Cytopenia of Undetermined Significance (ICUS) patients. Subsequently, we selected the optimal disease prediction model through performance comparison and conducted an analysis of relevant risk factors.

Results

We conducted an analysis of patients with cardiovascular disease, examining their medication history, test results, and medical records, and compared their characteristics with those who achieved the early LDL-C target. The results, including hazard ratios and cumulative incidence, clearly demonstrated the significant impact of early LDL-C goal attainment on reducing the recurrence rate of cardiovascular disease and a substantial decrease in healthcare resource utilization.

To predict disease progression in ICUS patients, we rigorously assessed three distinct models through 10-fold cross-validation. Furthermore, we integrated data from electronic medical records and evaluated the significance of clinical information embedded within textual data. Ultimately, the XGBoost (XGB) model, which incorporated text embedding data, exhibited the highest performance with an AUROC score of 0.817. Additionally, using Shapley values, we confirmed the meaningful contribution of textual data to the model's predictions.

Conclusions

First, we conducted a real-world analysis of disease prognosis and medical costs, aiming to provide valuable real-world evidence regarding the impact and benefits of early LDL-C reduction in Asian populations. This research can help inform treatment guidelines and support evidence-based medical decisions.

Second, we developed a machine learning model for predicting disease progression by leveraging unstructured clinical text data. By expanding our research to encompass diverse clinical information within electronic medical records, we aim to further contribute to supporting medical decisions and enhancing patient disease prognosis.

Keywords: electronic medical records, cardiovascular diseases, low-density lipoprotein cholesterol, major adverse cardiovascular events, artificial intelligence, machine learning, natural language processing

Contents

Abbreviation

ACAD: Asymptomatic coronary artery disease

ACR: Albumin-to-creatinine ratio

ACS: Acute coronary syndrome

AI: Artificial Intelligence

AMC: Asan Medical Center

ASCVD: atherosclerotic cardiovascular disease

AUROC: area under the receiver operating characteristic

BERT: bidirectional encoder representations from transformer

BMI: body mass index

CAG: coronary angiography

CCTA: coronary CT angiography

CI: confidence interval

CKD: chronic kidney diseases

CR: Coronary revascularization

CVD: cardiovascular disease

DBP: diastolic blood pressure

eGFR: estimated glomerular filtration

EHR: electronic health record

EMR: electronic medical record

FPR: false positive rate

HDL-C: high-density lipoprotein cholesterol

HF: heart failure

HR: hazard Ratio

HRU: healthcare resource utilization

ICUS: idiopathic cytopenia of undetermined significance

IS: ischemic stroke

LDL-C: low-density lipoprotein cholesterol

MACE: major adverse cardiovascular events

MDS: myelodysplastic syndrome

MI: myocardial infarction

ML: machine learning

MRI: magnetic resonance imaging

NLP: natural language processing

PAD: peripheral artery disease

PCI: Percutaneous coronary intervention

PCSK9: proprotein Convertase Subtilisin/Kexin Type 9

RAAS: Renin-angiotensin-aldosterone system

RCT: Randomized controlled trial

ROC: receiver operating characteristic

RWE: real-world evidence

SBP: systolic blood pressure

SVM: support vector machine

TIA: transient ischemic attack

TPR: true positive rate

UA: unstable angina

XGB: extreme gradient boosting

List of Tables

List of Figures

Introduction

Electronic medical records are a system that stores patients' medical information in digital form. Research using such medical data, combined with artificial intelligence technology, can demonstrate great potential in various fields such as disease prediction, treatment development, and medical decision support. there is. In particular, Real-World Evidence (RWE) research using electronic medical records can clearly identify relationships between diseases, drug effects, and side effects based on data obtained in real-world settings. This can help guide treatment and aid evidence-based medical decision-making.

The importance of early prevention and accurate treatment of cardiovascular disease (CVD), which causes significant mortality and health burden worldwide, is emphasized. Managing LDL-C levels has a significant impact on preventing the recurrence of cardiovascular disease, and recently, early treatment to reduce LDL-C levels has been recommended. However, because there is insufficient evidence on the effectiveness and benefits of actual treatment, we conducted a retrospective study using electronic medical records (EMR) to supplement this. Based on clinical evidence, the characteristics of the patient group were identified by defining the related diseases and medication history of patients with high-risk cardiovascular disease. Subsequently, statistical techniques were employed to compare the differences in cardiovascular disease recurrence rates and the frequency of medical resource utilization depending on whether the LDL-C goal was achieved early or not, and the results were verified. As a result, it was suggested that the findings of this study could serve as evidence to support medical decisions aimed at preventing the recurrence of cardiovascular disease.

Although patients' clinical information in electronic medical records, composed of free-form text, includes crucial medical data like test results and clinicians' opinions, it has been challenging to utilize it for research purposes due to limitations in processing methods. However, recent advancements in natural language processing technology have increased the potential for using medical data, leading to an expansion in clinical research applications. Building upon this progress, we defined a cohort of ICUS patients and developed a model to predict disease progression to myeloid malignancy by integrating structured and unstructured data. Subsequently, the presentation of key clinical variables provided insights into managing patients with unknown clinical characteristics, contributing to the improvement of disease prognosis.

In conclusion, we have presented concrete clinical evidence through a retrospective study using standardized EMR data, and we have enhanced the performance of the disease prediction model by incorporating text data from within the EMR. This can offer valuable insights for both patients and medical professionals to make informed treatment decisions and prevent diseases. Moreover, it is anticipated that our prediction model utilizing text data will contribute to clinical practice by providing more precise and dependable disease prediction results.

Chapter 1. MACE and direct medical costs in high-risk ASCVD patients reaching LDL-C targets sooner in South Korea (Early Reduction is Better)

Introduction

background

Cardiovascular disease (CVD) is recognized as a significant global issue related to mortality and public health.[1] The burden of cardiovascular disease has been exacerbated by the increase in major risk factors such as obesity, hypertension, and type 2 diabetes, as well as lifestyle factors including dietary habits and physical inactivity. In 1990, the total prevalence of CVD cases began at 271 million and nearly doubled to 523 million in 2019 (95% uncertainty interval [UI]: 497 million to 550 million).[2] Furthermore, the mortality rate associated with cardiovascular disease showed an approximately 1.5-fold increase, rising from 12.1 million (95% UI: 11.4 million to 12.6 million) in 1990 to 18.6 million (95% UI: 17.1 million to 19.7 million) in 2019.[2] Given that individuals who have experienced cardiovascular disease face an increased short-term risk of additional cardiovascular events, identifying and managing high-risk Atherosclerotic Cardiovascular Disease(ASCVD) patients is a clinically important challenge.[3]

Dyslipidemia is widely recognized as a major risk factor for cardiovascular disease (CVD) and stroke, and effective management of low-density lipoprotein cholesterol (LDL-C) levels is important, especially in high-risk patients. Past studies have repeatedly demonstrated that reducing LDL-C levels leads to a decreased risk of CVD and stroke.[4] According to a meta-analysis that included 25 statin trials, reducing LDL-C by 38.7 mg/dL resulted in a 23% reduction in the risk of experiencing Major Adverse Cardiovascular Events (MACE) compared to a placebo.[5] Building upon these research findings, the "Guidelines for the Management of Dyslipidemia and Prevention of Cardiovascular Disease," published by the American College of Clinical Endocrinologists (AACE) and the American College of Endocrinologists (ACE), have expanded the LDL-C target to <55 mg/dL, taking into account the patient's risk level.[6]

In particular, early reduction of LDL-C is recommended for individuals diagnosed with acute coronary syndromes, especially those considered to be at high risk for subsequent major adverse cardiovascular events (MACE). [7] Therefore, it is very important to apply effective treatments such as high-intensity statins early to patients hospitalized with acute coronary syndrome. [8] In a study of patients with acute myocardial infarction, a notable difference in the incidence of MACE was observed between the early and delayed LDL-C reduction groups. (3.4% vs. 9.4%; p=0.013) [9]

Emerging evidence underscores the clinical advantages associated with early reduction of low-density lipoprotein cholesterol (LDL-C) in randomized controlled trials (RCTs). However, it may be overly optimistic to expect uniform attainability of these outcomes in real-world clinical settings. These disparities arise due to the selective inclusion of specific patient cohorts in RCT studies. Therefore, it is imperative to bridge the evidence gap between RCTs and real-world clinical applications by conducting investigations that encompass real-world clinical evidence.[10] Research utilizing electronic health records (EMR) has proven particularly effective in providing substantial clinical evidence. EMR repositories contain a comprehensive spectrum of structured patient data, incorporating intricate aspects such as medication prescriptions and diagnostic records, complemented by unstructured patient data comprising a variety of test records. Recent findings from a nationwide Swedish cohort provide real-world clinical setting evidence for early LDL-C reduction. According to this, early reduction of LDL-C within 6 to 10 weeks after the onset of myocardial infarction reduced the risk of recurrent MACE or death, and the more substantial the degree of LDL-C reduction, the more pronounced the clinical benefit. [11]

A real-world evidence (RWE) study has also been conducted in Korea to investigate clinical outcomes following LDL-C treatment for cardiovascular disease prevention. A study involving patients with atherosclerotic cardiovascular disease (ASCVD) who underwent percutaneous coronary intervention (PCI) observed a correlation between reduced LDL-C levels and the incidence of Major Adverse Cardiovascular Events (MACE). However, this study was limited by its small sample size and a lack of systematic evaluation of lipid-lowering treatment patterns, imposing certain constraints.[12] In another study that explored the relationship between statin intensity and secondary prevention in 1,746 patients who achieved LDL-C treatment goals after PCI, the use of high-intensity statins was associated with a reduced risk of MACE compared to patients who did not use high-intensity statins, with statistically significant findings (4.1% vs. 9.9%; hazard ratio, 0.42; 95% CI, 0.23-0.79; $P<0.01$).[13] However, as this study was confined to patients who met LDL-C treatment goals, it raises uncertainty about its applicability to all ASCVD patients in Korea. Consequently, the establishment of a definitive LDL-C reduction target for preventing recurrent MACE in high-risk Korean ASCVD patients remains elusive, and subgroups at elevated risk have yet to be identified. Further clinical research in this domain is thus imperative, as it can contribute to narrowing the evidence gap between RCT and real-world clinical settings while enhancing our comprehension of the clinical significance of early LDL-C reduction in practical clinical contexts.

Objectives

In this study, we aimed to investigate the association between early LDL-C goal achievement and recurrent MACE among patients with ASCVD using electronic health records at a tertiary hospital in Korea and analyze the resulting differences in healthcare costs. Through this, we hope to leverage EMR to gain a clearer understanding of the importance and benefits of LDL-C management in real-world clinical settings.

Methods

Study Design

The study period was from January 2000 to December 2020, and study subjects were defined as patients admitted to the hospital with their first cardiovascular disease event between January 2000 and December 2019, considering at least 1 year of follow-up. The index date was defined as the time when LDL-C reassessment was recorded within 4 to 12 weeks after discharge. Patients were categorized based on their LDL-C test results at the time of LDL-C reassessment recording and followed until MACE recurrence or the end of the follow-up period. The overall study design is summarized in Figure 1.

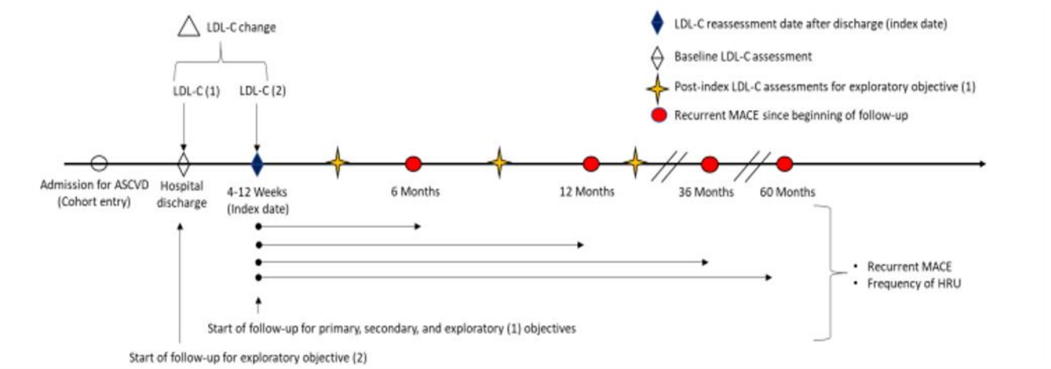

LDL-C change
LDL-C (1)
LDL-C (2)
LDL-C reassessment date after discharge (index date)
Baseline LDL-C assessment
Post-index LDL-C assessments for exploratory objective (1)
Recurrent MACE since beginning of follow-up
Admission for ASCVD (Cohort entry)
Hospital discharge
4-12 Weeks (Index date)
6 Months
12 Months
36 Months
60 Months
• Recurrent MACE
• Frequency of HRU
Start of follow-up for primary, secondary, and exploratory (1) objectives
Start of follow-up for exploratory objective (2)

Figure 1. Overall study design schema of recurrent MACE risk and HRU based on achievement of LDL-C goal in high-risk ASCVD patients. ASCVD: atherosclerotic cardiovascular disease; LDL-C: lipoprotein cholesterol; MACE: major adverse cardiovascular events; HRU: Healthcare resource utilization.

Data Extraction

EMR data was extracted using the ABLE (Asan Biomedical Research Environment) system, the EMR database of Asan Medical Center, Seoul, one of the largest tertiary care hospitals in Korea. ABLE is a data system in which patient information is de-identified so that only verified investigators can access and extract it. De-identification of data was performed in line with the health insurance portability and accountability act for Korea.[14] Access to extracted data was secured under the oversight of an Institutional Review Board (IRB). We utilized this to extract patient data, including diagnosis, laboratory, and reports, for ASCVD patients admitted to Asan Medical Center from January 1, 2000 to December 30, 2020.

Study population

We conducted a retrospective cohort study with patients admitted for first or recurrent ASCVD from January 2000 to December 2019. Cohort entry date was defined as the date of admission for ASCVD. If patients were hospitalized more than once during the study time period, only the first hospitalization was included. Patients who had LDL-C tests conducted during this ASCVD hospitalization period were included in the study population.

Data Preprocessing

Diagnoses of ASCVD events and comorbidities used in the study were confirmed through International Classification of Diseases, 10th version (ICD-10) coding to ensure the accuracy of the results. Additionally, relevant laboratory results and reports were considered when necessary. If a patient's discharge date from the emergency room coincided with the admission date, it was treated as a single hospitalization due to the same cause and counted as one hospitalization period.

CAG/CCTA test results were used to define patients with 50% or greater stenosis within the category of ACAD patients in ASCVD. Regular expressions were employed to extract CAG/CCTA test result values for processing unstructured text data. Initially, only results containing major epicardial vessels such as LM (Left main), LAD (Left anterior descending), LCX (Left circumflex), RCA (Right coronary artery), and RI (Ramus intermedius) were separated from the test results. Subsequently, in the case of CAG test results where the outcomes were expressed as numeric values (%), only the numeric values of the relevant items were extracted. For CCTA test results displaying outcomes as text values, only the text values of the items were extracted. Patients were classified as having 50% or greater stenosis if there was a value exceeding 50 among the CAG test results or if there were values marked as moderate or severe in the CCTA test results.

LDL-cholesterol values were used, both directly measured and estimated using the Friedwald equation: LDL-C = (total cholesterol - HDL-C) - triglycerides/5. For each test item, only values within clinically meaningful ranges were considered for each test item. These ranges were as follows: LDL-C (10 or more and 600 or less), Total cholesterol (30 or more and 600 or less), HDL-C (1 or more and 400 or less), Triglycerides (10 or more and 400 or less). The criteria for outliers were also applied to values obtained through the Friedwald equation. When LDL-C test values and Friedwald equation values were available on the same date, the directly measured LDL-C value took precedence. In cases where there were multiple LDL-C measurements on the same day, the minimum value among them was considered for analysis.

In order to define Lacunar infarction as one of the criteria for patient exclusion, we utilized information documented in the patient's stroke notes. Since this data is in an unstructured text format, we examined the presence of specific words used to identify Lacunar infarction. If any of the words "Lacunar," "Small vessel," or "SVD" were found in the entire text of notes related to stroke, we defined it as Lacunar infarction and excluded the respective patient from the study cohort.

The use of statins and other medications among the entire study population was considered based on prescriptions issued within one year of the index date. Statin use was categorized into high intensity, moderate intensity, and low intensity based on dosage, with dosage information extracted from prescription code names used within the AMC. In cases where multiple statin prescriptions occurred during the study period, only the prescription closest to the index date was taken into account. Furthermore, in instances where multiple statin prescriptions were given on the same date, only the prescription with the highest dosage was included.

The definition of "cardiac enzyme" includes patients who meet all of the following conditions. These cardiac enzymes were used in the definition of the outcome variable and were not considered as primary outcome. First, patients who received Troponin-I or CK-MB tests from the time of emergency room admission or hospitalization until before undergoing CAG, PCI, or CABG tests are included. The upper limit for Troponin-I test is defined as 1.5, and for CK-MB test, it is defined as 5. The units of measurement for both test results are standardized to ng/mL. Patients are included if their Troponin-I or CK-MB test results exceed the upper limit for the respective test. In cases where CAG, PCI, or CABG tests were performed multiple times on the same visit date, the earliest test performed in such situations was considered.

Key variables

In this study, the cohort entry date was defined as the date of admission for ASCVD. At this time, ASCVD (MI or Unstable Angina, Stable Angina, Asymptomatic CAD, IS/TIA, PAD) was defined as follows. For hospitalizations related to ASCVD, only records within approximately 30 days before or after the initial diagnosis registration were included in the analysis and CAG and PCI tests refer to records of tests conducted within 7 days before or after the diagnosis date. MI was defined as hospitalized patients diagnosed with I21-I23 codes and those who underwent CAG or PCI tests. Unstable Angina was defined as hospitalized patients diagnosed with I20 codes and those who underwent CAG or PCI tests. In this study, patients diagnosed with MI or Unstable Angina were categorized as the ACS group. Stable Angina was defined as hospitalized patients diagnosed with I20.8, I20.9, I24, I25.2-I25.5, I25.8-I25.9 codes and those who underwent CAG or PCI tests. ACAD was defined as patients diagnosed with I25.0, I25.1, or I25.6 codes and included patients who exhibited 50% or greater stenosis or moderate to severe findings in CAG or CCTA tests. IS/TIA was defined as patients diagnosed with I63 or G45.9 codes, and it included patients who had one or more brain CT or MRI imaging results within 30 days before or after the diagnosis code registration date. PAD was defined as patients diagnosed with I70, I73, I74 codes and included those who underwent coronary thrombosis and thrombosis-related surgeries, peripheral vascular surgeries, or vascular angiography within 7 days before or after the diagnosis code registration date. If a patient received multiple diagnoses related to ASCVD on the same day, priority was given in the following order: ACS, Stable Angina, and ACAD. Details regarding the definition of ASCVD within the inclusion criteria are summarized in eTable 1.

The definitions of variables describing the basic characteristics of patients are as follows: For all variables assessed multiple times, the value closest to the cohort entry date or index date was used. Age, gender, and smoking status were considered in relation to the cohort entry date. BMI, blood pressure values, eGFR, ACR, HDL-C, total cholesterol, triglycerides, and lipoprotein(a) test results considered data from the year prior to the index date. For BMI, values below the 25th percentile and above the 75th percentile were excluded. Only values within clinically meaningful ranges were considered for each test item: eGFR (1 or more and less than 200), ACR (30 or more and 1000 or less), HDL-C (1 or more and 400 or less), total cholesterol (30 or more and 600 or less), and triglycerides (10 or more and 3000 or less). Lipoprotein(a) used the entire range of results without excluding any values.

The definition of comorbidities in patients is as follows: This definition includes cases where patients have received a diagnosis at least once before the index date, starting from January 1, 2000. Chronic Kidney Disease (CKD) was defined as patients who had been diagnosed with the N18 code or patients with an eGFR value between the cohort entry date and index date that was less than 90. Diabetes mellitus included patients who had been diagnosed with E10-E14 codes or patients with an HbA1c value on the cohort entry date equal to or greater than 6.5%. Metabolic Syndrome was defined as including individuals who had met two or more of the following four criteria: 1) Serum triglyceride concentration equal to or greater than 150, 2) For males, HDL level less than 40; for females, HDL level less than 50, 3) Systolic blood pressure (SBP) equal to or greater than 130, or diastolic blood pressure (DBP) equal to or greater than 80, and 4) Fasting glucose equal to or greater than 100. Hypertension included patients who had been diagnosed with I10-I13, I15 codes or patients who had used at least one of the following: beta blockers, RAAS inhibitors, or calcium channel blockers. Congestive heart failure comprised patients who had been diagnosed with I42, I43, I50 codes, while Atrial fibrillation disease covered patients who had been diagnosed with the I48 code. Additionally, Cancer included patients who had been diagnosed with C00–C97 codes, and Inflammatory Disease contained conditions like rheumatoid arthritis, psoriasis, and HIV. Rheumatoid arthritis was defined by the M05, M06 codes, psoriasis by the L40 code, and HIV by the B20-B24 codes. Details regarding the definition of comorbidities summarized in eTable 2.

The definition of medication in patients is as follows: This definition includes the medication history of patients from their index date up to one year before. It contains the use of Statin, Ezetimibe, Fibrate, Niacin, Cholestyramine, PCSK9 inhibitor, Aspirin or P2Y12 inhibitor, Beta blocker, RAAS inhibitor, and Calcium channel blocker. rug data is based on AMC prescription codes and drug ingredient names, which are validated through clinical discussions. Detailed information regarding the components of each medication can be found in eTable 3.

Outcomes

The primary endpoint is defined as a composite outcome that includes the following five results: myocardial infarction (MI), ischemic stroke, hospitalization due to unstable angina, coronary revascularization (CABG or PCI), and all-cause mortality. The secondary endpoint is defined as a composite outcome that includes the following three results: myocardial infarction (MI), ischemic stroke, and all-cause mortality. All endpoint criteria were considered from the index date until the end of the study period on December 31, 2020. However, in cases where follow-up was terminated earlier due to the occurrence of MACE or loss to follow-up, events up to that point were taken into account. When a patient's outcomes were recorded multiple times, the earliest recorded date was used as the final outcome date. MI is defined by including I21-I23 codes and is limited to patients who had cardiac enzyme measurements during hospitalization and had records of CAG, PCI, or CABG procedures. Ischemic stroke includes I63 and G45.9 codes and is limited to patients with CT or MR imaging results. Hospitalization for unstable angina includes I20, I24.0, and I24.9 codes or patients defined as such if they had lower maximum values of Troponin-I and CK-MB test results during outcome events than the upper reference limit and had CT or MR imaging results within 30 days after the index date, as well as records of CAG, PCI, or CABG procedures, and had cardiac enzyme measurements during hospitalization. However, patients with the I25 code were excluded. CR is defined as patients with results of PCI or CABG procedures. All-cause mortality is defined as patients with documented in-hospital death dates or cancer-related death dates in their hospital records. The Secondary Endpoint is defined in the same way as the Primary Endpoint. Detailed definitions can be found in eTable 4.

The analysis of patient's cardiovascular-related healthcare utilization frequency includes the following seven variables. All these items are considered from the index date onwards until the most recent discharge date recorded for each patient in the EMR database. CV-related hospitalizations encompass the following conditions: 1) Hospitalization for unstable angina as defined in the Primary Endpoint, 2) Hospitalization for PCI or CABG surgery that does not overlap with condition 1, 3) Hospitalization for ischemic stroke (IS) as defined in the Primary Endpoint, 4) Hospitalization for reasons other than CABG in thoracic surgery, 5) Hospitalization for device insertion or arrhythmia procedures, 6) Hospitalization for cardiac examinations (CAG), cerebrovascular, or peripheral procedure, 7) Other CV/CS-related hospitalizations with a primary diagnosis related to circulatory system disorders. The length of stay for cardiovascular-related hospitalizations is defined by the duration of the respective hospitalization. Cardiovascular-related procedures include CABG, PCI, PTCA, and peripheral vascular procedures. Even if multiple tests were conducted on the same day, each instance is counted separately. Lipid panel tests encompass cases where Total Cholesterol, Triglycerides, HDL-C, and LDL-C tests were conducted. CV-related rehabilitation visits refer to cases where rehabilitation programs were conducted in an outpatient setting. Emergency room visits were extracted based on treatment type codes, specifically focusing on visits to the emergency room only. Annual all-cause outpatient visits were defined based on treatment type codes, focusing solely on outpatient visits.

Statistical Analysis

Participants were followed from January 1, 2000 until the date of a ASCVD event, or December 31, 2020, whichever came first. To determine the follow-up period for each patient, we calculated individual follow-up end dates. If an outcome event occurred after the index date, the follow-up end date was set as the date of that outcome event. In cases where no outcome was recorded, the follow up end date was determined as the date of the patient’s last recorded discharge in the EMR database. If there were no hospital records for a patient after the index date, the follow-up end date was set as the index date itself.

The cumulative incidence was calculated as the number of recurrent MACE events among high-risk ASCVD patients during the study period, divided by the number of patients at risk within the cohort. It represented the proportion of recurrent MACE events among high-risk ASCVD patients during the follow-up. Cumulative incidence curves were generated using the Kaplan-Meier approach, illustrating the time to recurrent MACE events stratified by LDL-C reduction groups during the follow-up (Kaplan-Meier estimates with 95% pointwise confidence intervals). Statistical significance was assessed using a log-rank test.

Crude incidence rates were calculated by dividing the number of outcomes by the sum of person-times, where observation time was adjusted for the end date of follow-up for each patient. It represented the proportion of recurrent MACE events among high-risk ASCVD patients during the follow-up. Cumulative incidence curves were generated using the Kaplan-Meier approach, illustrating the time to recurrent MACE events stratified by LDL-C reduction groups during the follow-up.

Cox proportional hazards models were used to evaluate the association between achieving early LDL-C goals and reducing the risk of recurrent MACE in the overall cohort. This model was used to calculate adjusted hazard ratios (HRs) by adjusting for covariates. Covariates considered included age (continuous, years), sex (categorical, male, female), smoking (categorical, never, past and current smoker), diabetes (categorical, 0, 1), and aspirin or P2Y12 inhibition (categorical). Brother, 0) was included. ,1), beta blockers (categorical, 0,1), RAAS inhibitors (categorical, 0,1), CKD (categorical, 0,1,), statin use (categorical, high, medium, low), and non-statin lipid-lowering therapy (Ezetimibe, Fibrate; categorical 0, 1). Participants with missing data on covariates were excluded from the final study population prior to analysis.

Healthcare resource utilization (HRU) during the follow-up period were assessed to account for the varying observation periods among study patients. The average number of HRU occurrences within 1 year after the start of follow-up and the total number of HRU occurrences during each patient's entire follow-up period were divided by the patient's follow-up duration to calculate the annual HRU per patient. The HRU average rates, calculated based on the number of events, were evaluated using negative binomial regression when the data exhibited overdispersion, and Poisson regression was used when overdispersion was not present.

We defined 2-sided P values of <0.05 as statistically significant. All data collection and statistical analyzes were performed using R software version 4.2.3 (R Foundation for Statistical Computing).

Results

Overall Cohort

A total of 53,440 patients registered at Asan Medical Center between January 1, 2000, and December 31, 2019, who had a clinically evident high-risk ASCVD history and had LDL-C measurement results during their hospitalization, were included. Patients without a follow-up LDL-C measurement within 4-12 weeks after discharge were excluded (n=30,412). In cases where multiple LDL-C measurements were conducted, the measurement closest to 8 weeks post-discharge was used, and in instances of multiple measurements on the same day, the minimum value was considered. This LDL-C reassessment date was defined as the study's index date. Patients with a history of lacunar infarction stroke or hemorrhagic stroke before screening (n=1,500), those who exceeded blood pressure criteria (SBP > 180 mmHg, DBP > 110 mmHg) at the time of ASCVD admission or LDL-C reassessment (n=1,014), individuals who experienced MI or non-hemorrhagic stroke within 4 weeks after ASCVD discharge (n=1,742), and those with statin medication duration of less than 2 weeks from ASCVD admission until the index date (n=1,838) were excluded from the analysis. Following these exclusion criteria, the final study cohort consisted of 16,934 patients, including those with ACS (n=4,168), Stable Angina (n=7,024), ACAD (n=2,193), IS/TIA (n=2,977), and PAD (n=185). Subsequently, to analyze the risk of recurrent MACE according to early LDL-C goal achievement, the entire study population was classified into the early LDL-C goal achievers group ("early achievers"), consisting of individuals with reassessment LDL-C measurements at or below 55 (mg/dL), and the non-achievers of the early LDL-C goal("non-achievers"), comprising those with measurements above 55 (mg/dL) threshold. The process of cohort formation is summarized in Figure 2.

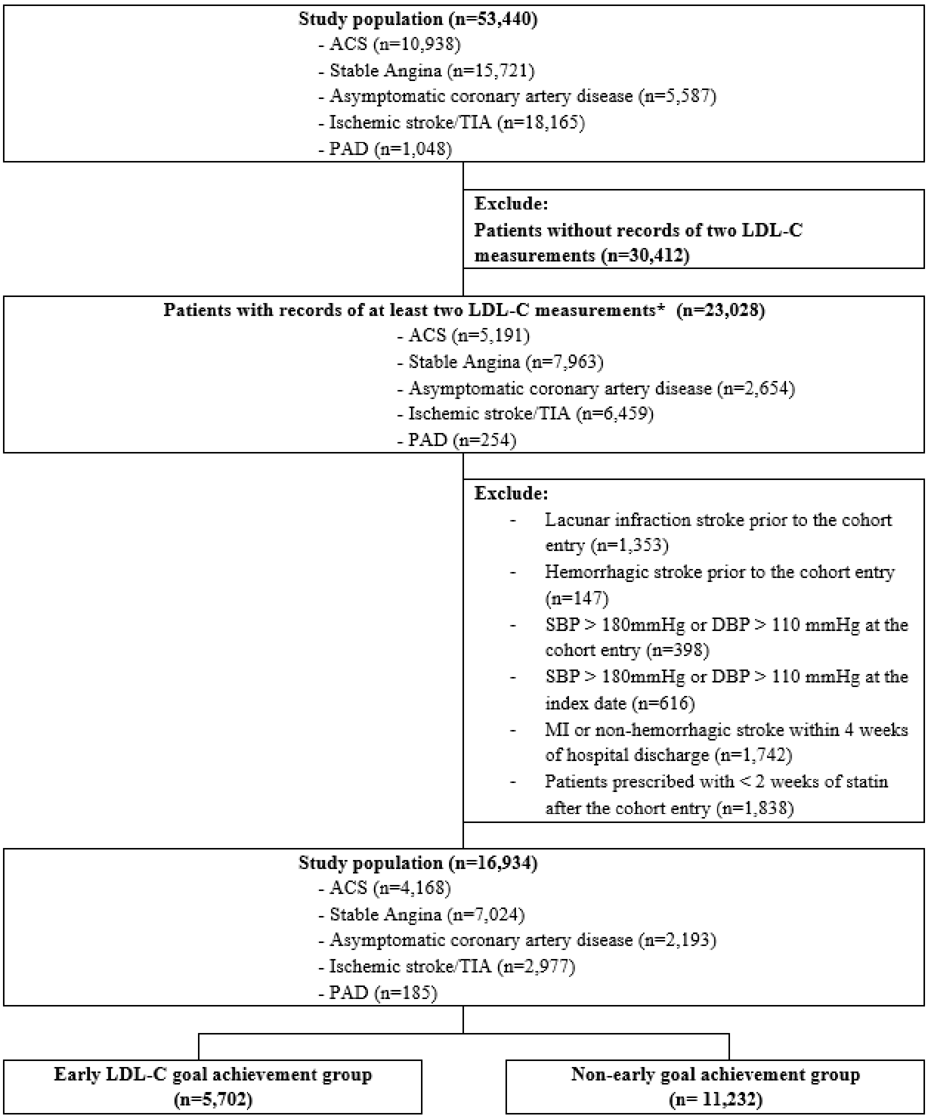

Study population (n=53,440)
- ACS (n=10,938)
- Stable Angina (n=15,721)
- Asymptomatic coronary artery disease (n=5,587)
- Ischemic stroke/TIA (n=18,165)
- PAD (n=1,048)
Exclude:
Patients without records of two LDL-C measurements (n=30,412)
Patients with records of at least two LDL-C measurements* (n=23,028)
- ACS (n=5,191)
- Stable Angina (n=7,963)
- Asymptomatic coronary artery disease (n=2,654)
- Ischemic stroke/TIA (n=6,459)
- PAD (n=254)
Exclude:
- Lacunar infraction stroke prior to the cohort entry (n=1,353)
- Hemorrhagic stroke prior to the cohort entry (n=147)
- SBP > 180mmHg or DBP > 110 mmHg at the cohort entry (n=398)
- SBP > 180mmHg or DBP > 110 mmHg at the index date (n=616)
- MI or non-hemorrhagic stroke within 4 weeks of hospital discharge (n=1,742)
- Patients prescribed with < 2 weeks of statin after the cohort entry (n=1,838)
Study population (n=16,934)
- ACS (n=4,168)
- Stable Angina (n=7,024)
- Asymptomatic coronary artery disease (n=2,193)
- Ischemic stroke/TIA (n=2,977)
- PAD (n=185)
Early LDL-C goal achievement group (n=5,702)
Non-early goal achievement group (n= 11,232)

Figure 2. Selection of the overall study cohort. ACS: acute coronary syndrome; DBP: diastolic blood pressure; LDL-C: low-density lipoprotein cholesterol; MI: myocardial infarction PAD: peripheral artery disease; SBP: systolic blood pressure; TIA: transient ischemic attack;

Baseline Characteristics

Baseline characteristics of the population are summarized in Table 1. The early-achievers counted 5,702 people, and the non-achievers counted 11,232 people. Angina history was more common in the early-achievers than in the non-achievers (46.6% versus 38.8%), but IS/TIA history was less common (14.6% versus 19.0%). The early-achievers had a higher proportion of men (76.5% vs. 68.3%), a higher prevalence of diabetes (38.5% vs. 32.0%), and a higher proportion on ezetimibe (32.4% vs. 17.1%). Most patients in both groups were receiving moderate-intensity statin treatment at the index date. (63.3% versus 70.2%).

	Early achiever [<55mg (n=5,702)]	Non-achiever (n=11,232)
	N (%)	N (%)
Mean (SD) Age (years)	63.541(10.89)	63.877(10.61)
Age ≥65 y (%)	2741(48.0)	5569(49.5)
Male	4366(76.5)	7678(68.3)
ASCVD subtype		
Angina	2661(46.6)	4363(38.8)
Asymptomatic CAD	698(12.2)	1495(13.3)
MI	1369(24.0)	2799(24.9)
Ischemic stroke/TIA	836(14.6)	2141(19.0)
PAD	40(0.7)	145(1.2)
Prior PCI/CABG	3931(68.9)	7003(62.3)
Smoking status		
Never smoker	1903(33.3)	4151(36.9)
Ex-smoker	177(3.1)	193(1.7)
Current smoker	1363(23.9)	2573(22.9)
Mean (SD) BMI(kg/㎡)	24.81(3.06)	24.88(3.117)
Mean (SD) SBP (mmHg)	128.43(19.86)	128.08(19.94)
Mean (SD) DBP (mmHg)	74.25(12.41)	74.209(12.40)
eGFR < 60 mL/min/1.73 ㎡	141(2.4)	174(1.5)
ACR ≥30mg/g	331(5.8)	563(5.0)
Level of stenosis		
Moderate stenosis (50-69%)	478(8.3)	680(6.0)
Severe stenosis (≥70%)	1878(32.9)	2673(24.0)
Chronic Kidney Disease	1751(30.7)	3421(30.4)
Diabetes mellitus	2196(38.5)	3604(32.0)
Metabolic syndrome	5543(97.2)	10823(96.3)
Hypertension	5390(94.5)	10429(92.8)
Congestive heart failure	145(2.5)	431(3.8)
Atrial fibrillation disease	283(4.9)	622.(5.5)
Cancer	433(7.5)	769(6.8)
Inflammatory disease	29(0.5)	61(0.5)
Rheumatoid arthritis	19(0.3)	39(0.3)
Psoriasis	10(0.1)	22(0.1)

HIV	0(0)	1(0)
Lipid lowering treatments		
Statin use	4426(77.6)	9258(82.4)
High intensity	796(17.9)	1142(12.3)
Moderate intensity	3608(81.5)	7880(85.1)
Low intensity	22(0.4)	236(2.5)
Ezetimibe	1853(32.4)	1921(17.1)
Fibrate	82(1.4)	277(2.4)
Niacin	2(0)	11(0.9)
Cholestyramine	0(0)	1(0)
PCSK9 inhibitors	0(0)	0(0)
Aspirin	5603(98.2)	10808(96.2)
P2Y12 inhibitor	5281(92.6)	9784(87.1)
Aspirin or P2Y12 inhibitor	5635(98.8)	10898(97.0)
Beta-blocker	3408(59.7)	6434(57.2)
RAAS inhibitor (ACE inhibitor, ARB, or aldosterone antagonist)	2522(44.2)	5302(47.2)
Calcium channel blocker	4462(78.2)	8475(75.4)
Baseline lipid profile		
Mean (SD) Total cholesterol (mg/dL)	160.33(41.36)	182.62(43.83)
Median (Q1-Q3) Total cholesterol (mg/dL)	158.0(129-188)	180(151-210)
Mean (SD) HDL-C (mg/dL)	43.89(11.82)	45.32(12.27)
Median (Q1-Q3) HDL-C (mg/dL)	42.0(36-51)	44.0(37-52)
Mean (SD) Triglycerides (mg/dL)	148.09(105.45)	146.04(92.32)
Median (Q1-Q3) Triglycerides (mg/dL)	123.0(87-178)	124.0(90-175)
Mean (SD) Lp(a) (mg/dL)	21.854(20.32)	30.942(30.10)
Median (Q1-Q3) Lp(a) (mg/dL)	15.0(7.9–28.2)	20.7(10.6-40.7)
LDL-C screening test		
Before cohort entry	2989(52.4)	5488(48.8)
After cohort entry	5450(95.5)	10530(93.7)
Mean (SD) LDL-C screening test frequency	22.38(23.58)	24.15(25.29)
Mean (SD) After cohort entry	17.33(17.79)	18.85(19.0)

Table 1. Baseline characteristics of the Study Population. ASCVD: atherosclerotic cardiovascular disease; LDL-C: low-density lipoprotein cholesterol; CAD: coronary artery disease; TIA: transient ischemic attack; PAD: peripheral artery disease; BMI: body mass index; SBP: systolic blood pressure; DBP: diastolic blood pressure; eGFR: estimated glomerular filtration rate; ACR: Albumin-to-creatinine ratio; PCSK9: proprotein Convertase Subtilisin/Kexin Type 9; HDL-C: high-density lipoprotein cholesterol; RAAS: Renin-

angiotensin-aldosterone system; ACE: angiotensin-converting enzyme; ARB: angiotensin receptor blocker;

Statistical Analysis
Outcomes

The results of the primary, secondary and individual endpoint for recurrent MACE are presented in Table 2. During the follow-up period, a total of 3,511 primary endpoint events occurred, with 986 (17.2%) in the group that achieved the early LDL-C goal and 2,525 (22.4%) in the group that did not achieve the early LDL-C goal. Patients who achieved the early LDL-C goal had a lower risk of developing cardiovascular events compared to patients who did not achieve the LDL-C goal. (crude incidence rate per 100 person-years, 3.81 versus 4.34; adjusted HR 0.89 [95% CI 0.82-0.96]).

During the follow-up period, a total of 2,545 secondary endpoint events occurred, including 706 (12.3%) and 1,839 (16.3%) in the early-achievers and the non-achievers, respectively. Compared with the non-achievers, the early achievers had a lower risk of cardiovascular disease. (crude incidence rate per 100 person year, 2.56 versus 2.92; adjusted HR 0.91 [95% CI 0.83-0.99])

Early achievers had lower primary incidence rates for all individual MACE endpoints and significantly lower adjusted HRs for hospitalizations due to unstable angina and coronary revascularization compared to those who did not achieve the early LDL-C goal.

Outcomes	**No. of events**	**Person years**	**Crude incidence rate per 100 PY**	**Crude HR (95% CI)**	**Adjusted HR* (95% CI)**
Primary endpoint					
Early achievers	986	25822.97	3.81	0.85 (0.79, 0.91)	0.89 (0.82, 0.96)
Non-achievers	2525	58068.39	4.34	1.00 (Reference)	1.00 (Reference)
Secondary endpoint					
Early achievers	706	27501.85	2.56	0.85 (0.78, 0.93)	0.91 (0.83, 0.99)
Non-achievers	1839	62960.10	2.92	1.00 (Reference)	1.00 (Reference)
Individual endpoints					
All-cause mortality					
Early achievers	262	29157.01	0.89	0.93 (0.8, 1.07)	0.9 (0.78, 1.05)
Non-achievers	682	68195.03	1.00	1.00 (Reference)	1.00 (Reference)
Myocardial infarction					
Early achievers	91	28769.13	0.31	0.92 (0.72, 1.17)	0.84 (0.65, 1.07)

Non-achievers	253	67114.34	0.37	1.00 (Reference)	1.00 (Reference)
Hospitalization for unstable angina					
Early achievers	55	28894.52	0.19	0.72 (0.53, 0.97)	0.67 (0.49, 0.91)
Non-achievers	196	67239.71	0.29	1.00 (Reference)	1.00 (Reference)
Ischemic stroke					
Early achievers	408	28119.78	1.45	0.86 (0.77, 0.96)	0.95 (0.84, 1.07)
Non-achievers	1076	64844.57	1.65	1.00 (Reference)	1.00 (Reference)
Coronary revascularization					
Early achievers	375	27222.73	1.37	0.89 (0.79, 1)	0.86 (0.76, 0.98)
Non-achievers	954	62733.53	1.52	1.00 (Reference)	1.00 (Reference)

Table 2. Adjusted HR for Primary, Secondary and Individual Endpoint associated with early LDL-C goal achievement group by different exposure definition; Adjusted for age, sex, diabetes mellitus, smoking, aspirin or P2Y12 inhibition, betablockers, RAAS inhibitors, CKD, statin use (3 groups), and non-statin lipid-lowering treatment; Primary Endpoint : a composite of MI, ischemic stroke, hospitalization for unstable angina, coronary revascularization(CABG or PCI) and all-cause mortality; Secondary Endpoint : a composite of MI, ischemic stroke, and all-cause mortality.

Cumulative incidence of recurrent MACE

Figure 3 shows the cumulative incidence of the primary endpoint for the overall cohort, showing that incidence rates remained lower in early achievers than in non-achievers over long-term follow-up.

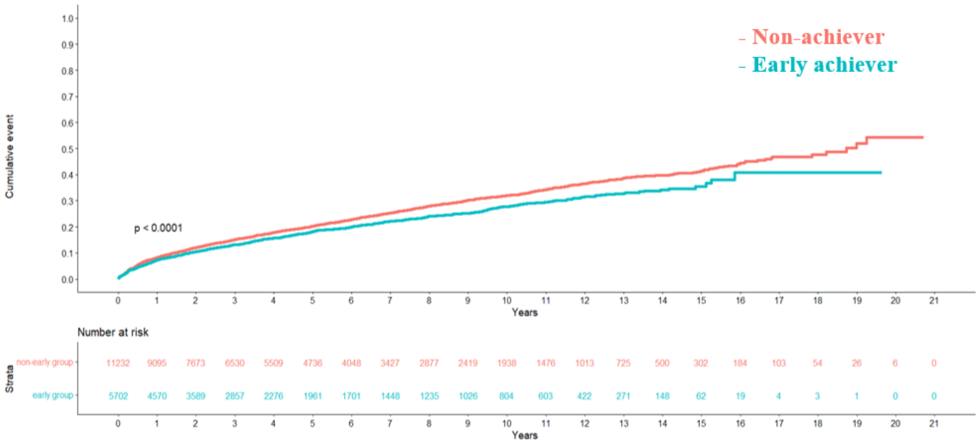


Figure 3. Cumulative incidence of recurrent MACE (primary composite MACE) among early achievers and non-achievers: a composite of MI, ischemic stroke, hospitalization for unstable angina, coronary revascularization (CABG or PCI) and all-cause mortality.

Subgroup analysis

The primary and secondary composite outcomes for recurrent MACE in ASCVD subgroups are summarized in Table 3. Achieving early LDL-C goals had the most significant impact on reducing the rate of cardiovascular disease recurrence in the ACS (MI or unstable angina) subgroup of patients. (Primary endpoint: crude incidence rate per 100 person-years, 3.10 versus 4.13, adjusted HR 0.73 [95% CI 0.63, 0.85]; secondary endpoint: crude incidence rate per 100 person-years, 1.98 versus 2.41, adjusted HR 0.82 [95% CI 0.68, 0.99]). For patients in the stable angina subgroup, a lower risk for the primary and secondary endpoints was observed in early achievers. (Primary endpoint: crude incidence rate per 100 person-years, 2.41 versus 2.85, adjusted HR 0.84 [95% CI 0.72, 0.97]).

LDL-C groups	No. of events	PY	Crude incidence rate (per 100 PY)	Crude HR (95% CI)	Adjusted HR[1] (95% CI)
ACS sub-group					
Primary endpoint					
Early achievers	244	7,846.78	3.109	0.74 (0.64, 0.86)	0.73 (0.63, 0.85)
Non-achievers	697	16,838.68	4.139	1.00 (Reference)	1.00 (Reference)
Secondary endpoint					
Early achievers	166	8,366.30	1.984	0.82 (0.69, 0.98)	0.82 (0.68, 0.99)
Non-achievers	454	18,788.90	2.416	1.00 (Reference)	1.00 (Reference)
Stable angina sub-group					
Primary endpoint					
Early achievers	268	11,082.63	2.418	0.83 (0.72, 0.96)	0.82 (0.68, 0.99)
Non-achievers	633	22,177.16	2.854	1.00 (Reference)	1.00 (Reference)
Secondary endpoint					
Early achievers	129	11,828.45	1.09	0.82 (0.66, 1)	0.81 (0.66, 1)
Non-achievers	329	24,129.30	1.363	1.00 (Reference)	1.00 (Reference)
Asymptomatic CAD sub-group					

Primary endpoint					
Early achievers	115	4,100.38	2.804	1.06 (0.85, 1.32)	1.04 (0.82, 1.31)
Non-achievers	242	9,170.61	2.638	1.00 (Reference)	1.00 (Reference)
Secondary endpoint					
Early achievers	66	4,403.15	1.498	1 (0.75, 1.34)	0.98 (0.72, 1.32)
Non-achievers	149	9,783.11	1.523	1.00 (Reference)	1.00 (Reference)

Table 3. Crude incidence rate and adjusted HR of MACE associated with LDL-C goal achievement among ASCVD subgroups; Adjusted for age, sex, diabetes mellitus, smoking, aspirin or P2Y12 inhibition, betablockers, RAAS inhibitors, CKD, statin use (3 groups), and non-statin lipid-lowering treatment;

Health resource utilization

The results of the analysis, which explored the relationship between achieving early LDL-C goals and healthcare resource utilization (HRU) in both the overall study population and within ASCVD subgroups, are summarized in Table 4. Across the entire cohort, individuals who achieved early LDL-C goals had lower rates of cardiovascular-related hospitalization (adjusted rate ratio 0.81 [95% CI 0.74, 0.89]) and a shorter length of stay for cardiovascular-related hospitalization compared to those who did not achieve these goals (adjusted RR 0.69 [95% CI 0.60, 0.78]).

These findings remained consistent within the ASCVD subgroups as well. In the ACS subgroup, early achievers experienced a 43% reduction in the length of stay for cardiovascular-related hospitalizations compared to non-achievers. (adjusted rate ratio 0.57 [95% CI 0.45, 0.71]) Similarly, in the Stable Angina subgroup, individuals who achieved early LDL-C goals also demonstrated a significant reduction in the length of stay for cardiovascular-related hospitalizations. (adjusted rate ratio 0.67 [95% CI 0.54, 0.82]) In the ACAD group, early-achievers demonstrated reduction in the length of stay for cardiovascular-related hospitalizations. (adjusted rate ratio 0.88 [95% CI 0.83, 0.92])

	HRU per patient year, mean (SD)*		
	Early achievers [<55mg/dL (n=5,702)]	Non-achievers (n=11,232)	Rate ratio (95% CI) **
Overall cohort			
CV-related hospitalization	0.45 (1.03)	0.48 (3.18)	0.81 (0.74, 0.89)

Length of stay of CV-related hospitalization	2.59 (9.9)	3.4 (24.11)	0.69 (0.6, 0.78)
CV-related procedure (CABG, PCI, PTCA, PAD-related procedures)	0.31 (0.53)	0.3 (0.59)	0.96 (0.85, 1.08)
Lipid panel tests	6.74 (9.32)	6.65 (8.38)	1.01 (0.98, 1.03)
CV-related rehabilitation visit	0.93 (0.74)	0.99 (2.07)	1.05 (0.93, 1.18)
Emergency room visit	0.68 (1.94)	0.58 (1.2)	1.03 (0.95, 1.13)
All-cause outpatient visits per year (including CV-related rehabilitation visit)	5.28 (11.52)	5.51 (11.66)	0.99 (0.96, 1.02)
ACS (MI/Unstable angina) sub-cohort			
CV-related hospitalization	0.43 (1.576)	0.393 (1.063)	0.92 (0.77, 1.09)
Length of stay of CV-related hospitalization	1.95 (8.462)	3.048(30.296)	0.57 (0.45, 0.71)
CV-related procedure (CABG, PCI, PTCA, PAD-related procedures)	0.285 (0.513)	0.297 (0.632)	0.86 (0.7, 1.06)
Lipid panel tests	6.749 (7.45)	6.528 (7.318)	1.06 (1.01, 1.12)
CV-related rehabilitation visit	0.904 (1.073)	1.176 (3.732)	1.33 (0.96, 1.84)
Emergency room visit	0.66 (1.87)	0.546 (1.219)	1.13 (0.97, 1.32)
All-cause outpatient visits per year (including CV-related rehabilitation visit)	4.751 (6.727)	5.189 (12.549)	0.93 (0.88, 0.99)
Stable angina sub-cohort			
CV-related hospitalization	0.451 (0.693)	0.371 (0.75)	0.98 (0.85, 1.13)
Length of stay of CV-related hospitalization	2.124 (9.889)	1.852 (10.536)	0.67 (0.54, 0.82)
CV-related procedure (CABG, PCI, PTCA, PAD-related procedures)	0.297 (0.426)	0.284 (0.505)	0.98 (0.81, 1.17)
Lipid panel tests	6.348 (7.165)	6.15 (7.995)	1.01 (0.97, 1.05)
CV-related rehabilitation visit	0.947 (0.63)	0.963 (1.534)	1.01 (0.88, 1.01)
Emergency room visit	0.575 (0.97)	0.46 (0.782)	1.02 (0.88, 1.19)
All-cause outpatient visits per year (including CV-related rehabilitation visit)	4.773 (11.613)	4.517 (5.337)	1.08 (1.04, 1.12)
Asymptomatic CAD sub-cohort			

CV-related hospitalization	0.389 (0.651)	0.467 (1.881)	0.98 (0.76, 1.25)
Length of stay of CV-related hospitalization	2.402 (8.429)	3.21 (26.038)	0.88 (0.83, 0.92)
CV-related procedure (CABG, PCI, PTCA, PAD-related procedures)	0.322 (0.613)	0.268 (0.336)	1.29 (0.99, 1.67)
Lipid panel tests	8.074 (17.12)	6.697 (7.97)	1.14 (1.06, 1.24)
CV-related rehabilitation visit	0.813 (0.575)	0.884 (0.606)	1.33 (0.84, 2.1)
Emergency room visit	0.51(0.784)	0.488 (0.862)	0.91 (0.71, 1.17)
All-cause outpatient visits per year (including CV-related rehabilitation visit)	5.974 (11.058)	5.348 (9.102)	1.09 (1.01, 1.18)

Table 4. Healthcare resource utilization among early achievers versus non-achievers per patient years.; adjusted for age, sex, diabetes mellitus, smoking, aspirin or P2Y12 inhibition, betablockers, RAAS inhibitors, Chronic Kidney Disease, statin use (3 groups), and non-statin lipid-lowering treatment

* HRU per patient years: HRU per patient year is calculated by the total HRU for each individual, divided by their total follow-up year, which would give a rate of HRU per follow-up year;

** If the data is over dispersed (e.g., variance of HRU is greater than the mean value), use negative binomial regression instead Poisson regression.

Discussions

In this study, we examined the impact of achieving the early LDL-C target on the recurrence of cardiovascular disease in high-risk ASCVD patients hospitalized at Asan Medical Center in Seoul. Our findings suggest that effective management of LDL-C levels can significantly decrease the incidence of recurrent MACE and the utilization of healthcare resources. Our study aims to bridge the gap between academic knowledge and real-world medical practice by highlighting the advantages of early goal achievement and providing insights into the necessary thresholds for its attainment.

Additionally, this study investigated the advantages of early achievement of LDL-C goals, with a specific focus on high-risk ASCVD patients overall and the ACS and stable angina subgroups. This comprehensive approach allowed us to examine the topic from various perspectives and gain a comprehensive understanding. In addition to analyzing the recurrence rates of MACE, we conducted a detailed examination of the frequency of medical resource utilization associated with early goal achievement to present comprehensive findings.

However, this study has certain limitations as it is a retrospective study conducted at a single institution. For instance, some patients may have been lost to follow-up due to transfers to other hospitals, potentially resulting in discrepancies in clinical characteristics between the excluded patients and the study group. These limitations could restrict the generalizability of our findings to the broader population of ASCVD patients in Korea. To address these limitations, we intend to expand the study using multicenter data in the future.

Furthermore, this study encountered limitations related to the use of EMR data. Much of the clinical information preserved in EMRs is documented in unstructured text format, making it difficult to analyze and extract meaningful insights. IS/TIA subgroups were defined based on free-form text records such as brain MR and CT, and information loss may occur during the patient group selection process due to limitations in text processing methods. The integration of advanced NLP techniques is necessary to harness the potential of unstructured text data and minimize information loss. As part of our future research efforts, we plan to leverage NLP techniques to conduct comprehensive EMR data analysis and AI-based investigations.

Conclusion

We conducted a study using diverse clinical information from Korea's hospital EMR database to examine the impact of lowering LDL-C levels to target values in high risk ASCVD patients on the recurrence rate of MACE and subsequent HRU. The findings of this study present real-world clinical evidence regarding the advantages of achieving early LDL-C targets in reducing the recurrence rate of cardiovascular disease.

Chapter 2. Predicting Disease Progression from Idiopathic Cytopenia of Undetermined Significance to Myeloid Malignancies using a Machine Learning-Based Approach

Introduction

Backgrounds

Idiopathic cytopenia of undetermined significance (ICUS) is diagnosed when unexplained persistent cytopenia persist for more than 6 months and bone marrow tests cannot provide a specific disease diagnosis. This is used as a diagnostic category for patients with cytopenia who do not meet criteria for myelodysplastic syndrome (MDS).[15] The annual probability of a patient with ICUS progressing to MDS is estimated to be approximately 0.5-1%, with a higher probability associated with the onset of MDS, specific somatic mutations, or evidence of clonal hematopoiesis associated with an increased number of genetic mutations and allele frequencies. However, not much research has been done on the impact of clinical variables other than genetic factors on disease progression or survival in ICUS.

In contrast, within the context of myelodysplastic syndrome (MDS), several well-established prognostic systems are in place to predict progression to acute myeloid leukemia (AML) and assess survival outcomes. These prognostic measures encompass the International Prognostic Scoring System (IPSS), revised IPSS, WHO Classification-based Prognostic Scoring System, and Lower-Risk Prognostic Scoring System. They take into account factors such as the severity of cytopenia, bone marrow blast percentages, bone marrow chromosomal findings, and patient age.[16] Given that a significant proportion of ICUS patients exhibit similar genetic abnormalities and clinical characteristics to those with MDS, the development of predictive tools tailored to ICUS is imperative.

Electronic medical records (EMRs) encompass a diverse range of clinical information, including diagnoses, medications, and test results, for extensive patient cohorts. Notably, advancements in artificial intelligence and machine learning technology have facilitated the identification of intricate relationships within EMR data, leading to effective research in disease progression prediction. [17,18] Nevertheless, EMRs comprise both structured and unstructured text data. Clinical text data, which often includes vital clinical information like medical records, test result reports, and genomic testing records, is typically presented in free-form notes, posing challenges for data processing. [19,20] Recent strides in natural language processing (NLP) have yielded methodologies capable of extracting meaningful clinical insights from narrative text data. Transformer-based models like BERT, which consider bidirectional context, have proven instrumental in enhancing the predictive performance of models leveraging textual data content.[21] A BERT-based model trained on authentic Electronic Health Records (EHRs) exhibited an over 8% performance improvement in predicting over 300 diseases.[22] Particularly noteworthy is the fact that critical variables influencing disease outcomes related to myeloid leukemia progression, such as cytogenetic information and bone marrow myeloblasts [23] are predominantly recorded in unstructured text data. Therefore, effective clinical text processing is imperative for ICUS disease prediction.

Objectives

In this study, we leveraged these natural language processing and machine learning techniques to develop and validate a disease prediction model for ICUS that leverages the rich clinical information from both structured EMR data and unstructured text data.

Methods

Study Design

The study period was from January 2000 to December 2021, and during this period, patients who underwent bone marrow examination at Asan Medical Center and met the ICUS diagnostic criteria were selected as study subjects. Subsequently, additional bone marrow examinations were conducted, and patients diagnosed with myeloid malignancies were categorized into the target patient group. To predict disease progression in patients, we selected an optimal machine learning model using a dataset comprising structured data features. Furthermore, we compared the model's performance on three datasets, including both structured and unstructured data. Additionally, we visualized the factors influencing the model.

Data Preprocessing

Structured data

We extracted structured data from the electronic medical record (EMR) system at Asan Medical Center in Seoul. This data included diagnoses, medications, physical information, laboratory test results, and the patient's Previous medical history. These data points were organized based on the date of the initial bone marrow examination, with priority given to records closest to that date.

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

Diagnosis table

The diagnoses utilized in this study were validated through KCD7 coding to ensure result accuracy. We considered all current and past comorbidities, with the current medical history focusing on conditions diagnosed between the date of admission and discharge, concurrent with the initial bone marrow examination. Furthermore, for past medical history, only records preceding the bone marrow examination were included in the analysis. Among these, they were incorporated into the final dataset in the order of common diagnoses in the entire patient group. Specifically, the current medical condition comprised a total of 10 characteristics, while the past medical condition comprised a total of 12 characteristics. Each disease was represented using one-hot encoding based on its presence or absence.

Medication table

Each patient's medication history consisted of a record of all medications prescribed before the first bone marrow examination. Among them, the top 37 drug ingredient names commonly prescribed in the entire patient group were included in the analysis. Each drug was represented using one-hot encoding based on whether it was taken or not.

Physical information table

The body information table included variables such as height, weight, BMI (body mass index), and BSA (body surface area). To control outliers, BMI exceeding 50 was designated as a missing value. If a measure did not exist, we replaced that value with -1. All physical information values are included as numerical values.

Laboratory test result table

The patient's laboratory test table comprised 33 unique lab test results after eliminating duplicates. Special characters within the lab test names were eliminated, and only the most frequently occurring test items were chosen as features. Furthermore, for cases where multiple tests were conducted on the same day, the median value was computed from the results and incorporated as numerical data.

Previous medical history

The previous medical history table encompasses records pertaining to the patient's historical medical conditions, family medical lineage, alcohol consumption patterns, and smoking tendencies. These variables within the previous medical history table are categorized as categorical and originate from the patient information questionnaire. In cases where these variables were documented in text format, we employed regular expressions to extract the relevant information, subsequently incorporating it into our analytical framework. Within the section dedicated to historical medical conditions, binary records (categorical, 1,0) were included to signify the presence or absence of diseases such as diabetes mellitus (DM), hypertension (HTN), tuberculosis (Tbc), and hepatitis. Similarly, in the family medical history section, binary indications (categorical, 1,0) were included to denote the presence or absence of cancer, diabetes, and hypertension. Concerning alcohol consumption habits, a value of 1 was assigned solely to indicate alcohol consumption, while abstention from drinking and complete non-drinking were regarded equivalently (categorical, 1,0). Similarly, smoking habits were coded in the same way (categorical, 1,0).

Unstructured data

Bone marrow test result

The bone marrow test result sheet serves as a comprehensive document encompassing the outcomes of the bone marrow examination, coupled with a diverse array of clinical insights. This document holds significant unstructured data, including intricate details of the test findings, numerical ratios for each cell type, and the expert opinions of clinicians. Therefore, the bone marrow test report plays a pivotal role in ensuring a precise comprehension and effective utilization of the test outcomes. In our research, we divided the bone marrow test results into two fundamental segments and conducted preprocessing procedures accordingly.

We extracted numerical percentage values from the bone marrow examination results. We employed regular expressions to systematically extract percentage values corresponding to different cell types. Subsequently, these extracted values were meticulously organized and stored in a database structured according to each characteristic.

Concurrently, we extracted clinicians' assessments from the test results. This specific section is typically presented in a free-form text format. As it will serve as input to the language model in subsequent steps, we opted for a straightforward approach that utilizes regular expressions to extract only the portions that correspond to the clinician's final conclusions without any additional preprocessing.

Chromosome test result

Chromosomal analysis is categorized as essential text that must be incorporated due to its clinical relevance to bone marrow cells. As it is embedded within the EMR in an unstructured free-text format, the utilization of this data necessitates a text preprocessing procedure. In this research, two analytical approaches have been selected to leverage the results of chromosomal analysis.

First, we integrated karyotype results as structured attributes. Karyotype encompasses details regarding the total chromosome count, sex chromosome composition, and descriptions of any observed chromosomal irregularities. In cases where karyotypes were presented in unstructured free-text format, we initiated a segmentation process utilizing appropriate delimiters. This was followed by the elimination of sex chromosome data and any extraneous special characters. Subsequently, with guidance from clinical experts, we selectively extracted symbols and numerical values that held significance. The respective definitions for each symbol are delineated as follows:

- Plus sign: Gain
- Minus sign: Loss
- del: Deletion or loss of chromosome material. May be either terminal or interstitial
- inv: Inversion of a chromosome segment: breakpoints may be on either side of the centromere (pericentric) or within the same chromosome arm (paracentric)
- t: Balanced translocation involving two or more chromosomes. Also use “t” to describe balanced whole-arm translocations. See text for reporting Robertsonian translocations.

Second, we initiated the process of extracting the clinician's evaluations from the test results. This specific section is typically presented in an unstructured text format. Given its intended use as input to the language model in conjunction with the subsequent bone marrow test results, we performed a straightforward preprocessing using regular expressions to selectively extract only the clinicians' ratings.

To process unstructured textual data, we employed two distinct methodologies. Initially, we employed a canonicalization approach using regular expressions, and subsequently, we utilized a text extraction method for the purpose of serving as input to the Pre-trained Language Model (PLM). To comprehensively encompass all clinical information recorded within these two distinct text sources, we amalgamated the chromosome text results and bone marrow examination conclusions into a unified text format. These text inputs were then transformed into 768-dimensional features through the Pretrained Language Model for integration into the model.

Predictive model

Learning Dataset

In order to assess model performance with respect to the incorporation of textual data, we partitioned the dataset into three distinct segments. Initially, we constructed a dataset comprising 110 standardized features, exclusively encompassing essential patient information. Subsequently, we curated a dataset consisting of 251 features, effectively integrating basic patient details with standardized text data. Lastly, we created a comprehensive dataset containing a total of 1,109 features. This comprehensive dataset includes fundamental patient information, standardized text data, and unstructured text features that have been embedded using a language model.

Natural Language Model

Our study leveraged pre-trained language models to seamlessly integrate unstructured text data from EMR into predictive models. BERT, based on the Transformer architecture, creates models with high prediction performance by demonstrating an excellent ability to effectively encode information within sentences. However, these models are typically pre-trained and evaluated using general domain text, which is problematic when evaluating their effectiveness on datasets consisting of biomedical text.[21] Recent advances have tried to alleviate these problems by exploring BERT-based models that undergo special pre-training to include terms widely used in the biomedical domain.[24,25] We leveraged the power of a domain-specific language model known as PubMedBERT for embedding extraction. PubMedBERT was pretrained from scratch based on abstracts retrieved from PubMed and full-text articles extracted from PubMedCentral. This study shows that PubMedBERT consistently outperforms BERT-based models on a variety of biomedical NLP tasks.[26]

We performed a careful fine-tuning procedure to adapt PubMedBERT to our unique dataset. In the initial stage, we preprocessed the vocabularies present in the bone marrow test results and chromosome text results to match the characteristics of each dataset and incorporated them into the model's lexicon. In particular, in the case of bone marrow test results, much of the personal information, including the clinician's name, was recorded in Korean, so a selective approach was applied to exclude non-English words. Most of the chromosome text results were written in Korean, and only words containing special characters were excluded. Classification of these words was performed systematically using regular expressions.

After adding words related to bone marrow test results to the model lexicon, we performed comprehensive pre-training of the model on the entire text corpus. Then, to derive the final layer output from the model, a 768-dimensional vector was generated using a mean pooling technique. These vectors were then seamlessly incorporated into the machine learning model.

Basic ML model

To identify the most suitable model, we conducted a performance evaluation by comparing the results obtained from the Extreme Gradient Boosting Model (XGB) [27], Support Vector Machine (SVM) [28], and Logistic Regression models.

Within the scope of this study, XGBoost (Extreme Gradient Boosting) emerged as the final choice. XGBoost is a machine learning methodology based on the gradient boosting algorithm, utilizing an ensemble learning technique that leverages the collective strength of multiple decision trees. The algorithm functions by sequentially learning new trees that compensate for the prediction errors of previous trees, continually improving the model's accuracy. XGBoost incorporates L1 and L2 regularization techniques to enhance accuracy by carefully constraining the complexity of the decision tree and mitigating the risk of overfitting. Furthermore, XGBoost excels at clearly identifying the most influential variables, thereby enhancing the model's interpretability. Additionally, its unique support for parallel processing ensures swift learning even when handling extensive datasets. Consequently, XGBoost excels at delivering consistent and well-generalized predictions, making it widely adopted in the field of machine learning.

Evaluations

Individuals with a subsequent diagnosis of myeloid malignancies in ICUS were assigned a positive label (1), and individuals without such diagnosis were assigned a negative label (0). To facilitate comparison and evaluation of candidate models, we used performance metrics including accuracy, sensitivity (positive recall), and false positive rate. Following the model training and validation steps, the area under the ROC curve (AUROC) was assessed to measure the performance of the final model.

To mitigate potential biases arising from unbalanced class distributions, all models were trained using Stratified K-fold cross-validation.[30] Layered K-fold cross-validation, which accounts for label distribution imbalance, entails dividing the entire data set into k subsets and using these systematically for testing purposes. The final performance indicator was calculated as the average of these 10 folds. In addition, we performed optimal hyperparameter tuning for each model through grid searching algorithm.

Results

Study populations

This study was conducted using a cohort of 26,393 patients who underwent bone marrow examinations at AMC between January 1, 2000, and December 31, 2021. However, patients who were diagnosed with a specific medical condition at any point (n=17,193) or those who did not diagnosed with cytopenia in bone marrow test (n=1,094) were excluded from the study. Cytopenia was defined as follows: hemoglobin <13 g/dL in males or <12 g/dL in females, neutrophil count <1.9 × 10^9/L, and/or platelets <150 × 10^9/L. However, patients who were diagnosed with a hematological disorder but did not undergo subsequent bone marrow testing were explicitly excluded (n=1,144).

Furthermore, individuals under the age of 18 at the time of cytopenia diagnosis (n=1,073) and those with cytopenia attributed to other causes (n=4,074) were also excluded. Cytopenia due to other causes encompassed patients with prior exposure to cytotoxic chemotherapy or radiation therapy (n=497), patients with a history of various conditions such autoimmune disorders, and transplantation (n=584), as well as patients with documented blood or bone marrow disorders before the initial testing (n=2,431). It also included patients who received a diagnosis of a blood disorder within 30 days of their first bone marrow test (n=226), individuals with a history of active infection by atypical pathogens in the month prior to cytopenia diagnosis (n=107), and patients who used immunosuppressant and chemotherapeutic agents in the month preceding their diagnosis (n=26). Additionally, those who were deemed ineligible after a clinician's chart review were excluded (n=203).

After applying these stringent exclusion criteria, the final study population that met the criteria for ICUS diagnosis comprised a total of 1,815 patients. Subsequently, patients were categorized into two groups: those who received a blood disorder diagnosis during additional bone marrow testing and those who did not undergo further testing or were not subsequently diagnosed with a blood disorder. This process is depicted in Figure 4.

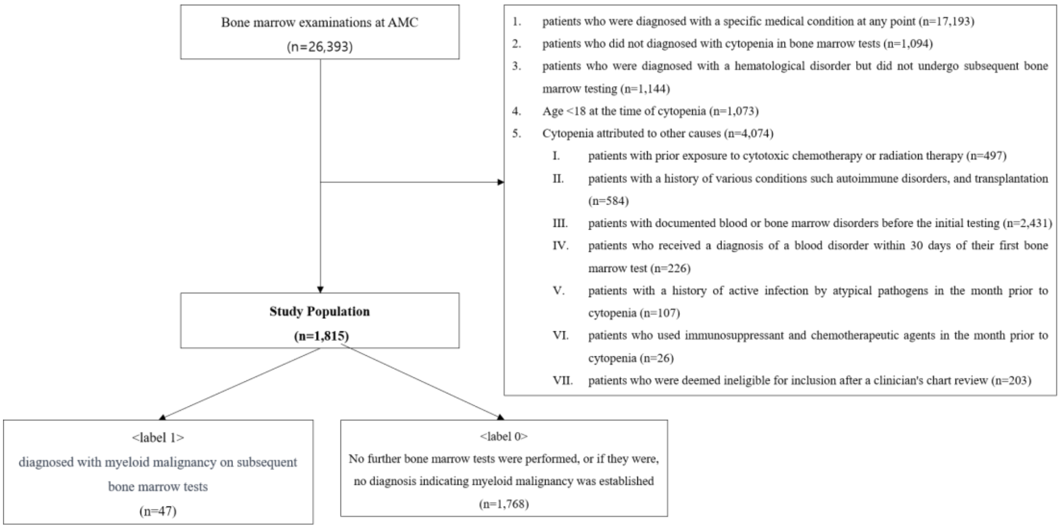

Bone marrow examinations at AMC
(n=26,393)
1. patients who were diagnosed with a specific medical condition at any point (n=17,193)
2. patients who did not diagnosed with cytopenia in bone marrow tests (n=1,094)
3. patients who were diagnosed with a hematological disorder but did not undergo subsequent bone marrow testing (n=1,144)
4. Age <18 at the time of cytopenia (n=1,073)
5. Cytopenia attributed to other causes (n=4,074)
I. patients with prior exposure to cytotoxic chemotherapy or radiation therapy (n=497)
II. patients with a history of various conditions such autoimmune disorders, and transplantation (n=584)
III. patients with documented blood or bone marrow disorders before the initial testing (n=2,431)
IV. patients who received a diagnosis of a blood disorder within 30 days of their first bone marrow test (n=226)
V. patients with a history of active infection by atypical pathogens in the month prior to cytopenia (n=107)
VI. patients who used immunosuppressant and chemotherapeutic agents in the month prior to cytopenia (n=26)
VII. patients who were deemed ineligible for inclusion after a clinician's chart review (n=203)
Study Population
(n=1,815)
<label 1>
diagnosed with myeloid malignancy on subsequent bone marrow tests
(n=47)
<label 0>
No further bone marrow tests were performed, or if they were, no diagnosis indicating myeloid malignancy was established
(n=1,768)

Figure 4. The flow chart of study populations; AMC: asan medical center;

Performance of the ML model

Our study utilizes three distinct datasets. The base dataset comprises structured patient information. The base+text dataset additionally incorporates text data structured using regular expressions. Finally, the base+text+embedding dataset consists of free-text data that has been embedded using a language model. The experiment proceeds as follows: Firstly, the optimal machine learning model is selected using the foundational dataset. Subsequently, the performance of the optimal model is compared across the three datasets.

Model performance evaluation was conducted employing stratified 10-fold cross-validation procedure, and the summarized AUROC scores are presented in Table 5. As a result of assessing prediction performance on the basic dataset, XGBoost (XGB) demonstrated the highest performance, achieving a score of 0.77. Furthermore, the performance differences across all folds were smaller compared to SVM and logistic regression. Consequently, XGBoost, which exhibited the most stable and high performance, was chosen as the final model.

Subsequently, we compared the performance of each XGBoost model using the three constructed datasets. XGBoost achieved a performance score of 0.77 in the basic dataset, 0.788 in the dataset incorporating structured text, and recorded the highest performance of 0.817 in the dataset containing embedded text data. This indicates the positive impact of textual clinical data on the enhancement of disease prediction model performance. The ROC curve illustrating the experimental results is presented in Figure 5.

Dataset	**base dataset**			**base+text dataset**	**base+text+embedding dataset**
Number of folds	**XGB**	**SVM**	**Logistic Regression**	**XGB**	**XGB**
1-fold	0.734	**0.526**	**0.589**	0.697	0.841
2-fold	0.800	0.666	0.691	0.776	0.648
3-fold	0.767	0.659	0.774	0.882	0.896
4-fold	0.714	0.658	0.713	0.868	0.897
5-fold	0.767	0.842	0.788	0.744	0.856
6-fold	0.871	**0.883**	**0.931**	0.898	0.905
7-fold	0.693	0.792	0.745	0.753	0.777
8-fold	0.793	0.576	0.757	0.755	0.792
9-fold	0.826	0.713	0.621	0.85	0.782
10-fold	0.741	0.740	0.788	0.656	0.776
Mean	**0.77**	0.706	0.75	**0.788**	**0.817**

Table 5. Evaluation by AUROC score of 10-fold cross-validation for each model. SVM: support vector machine; XGB: extreme gradient boosting;

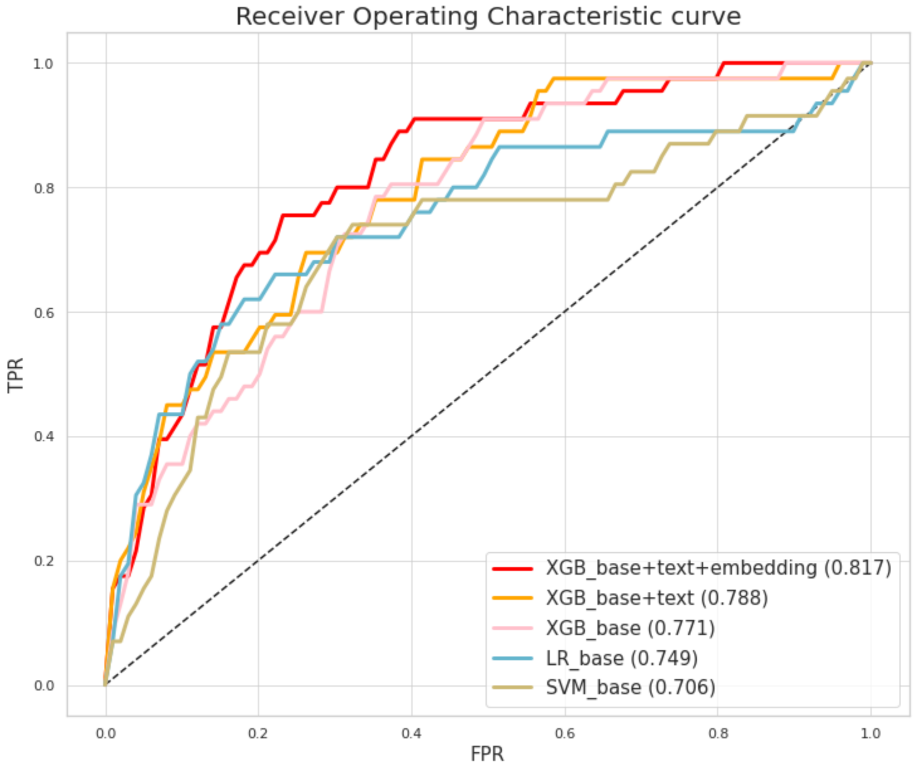


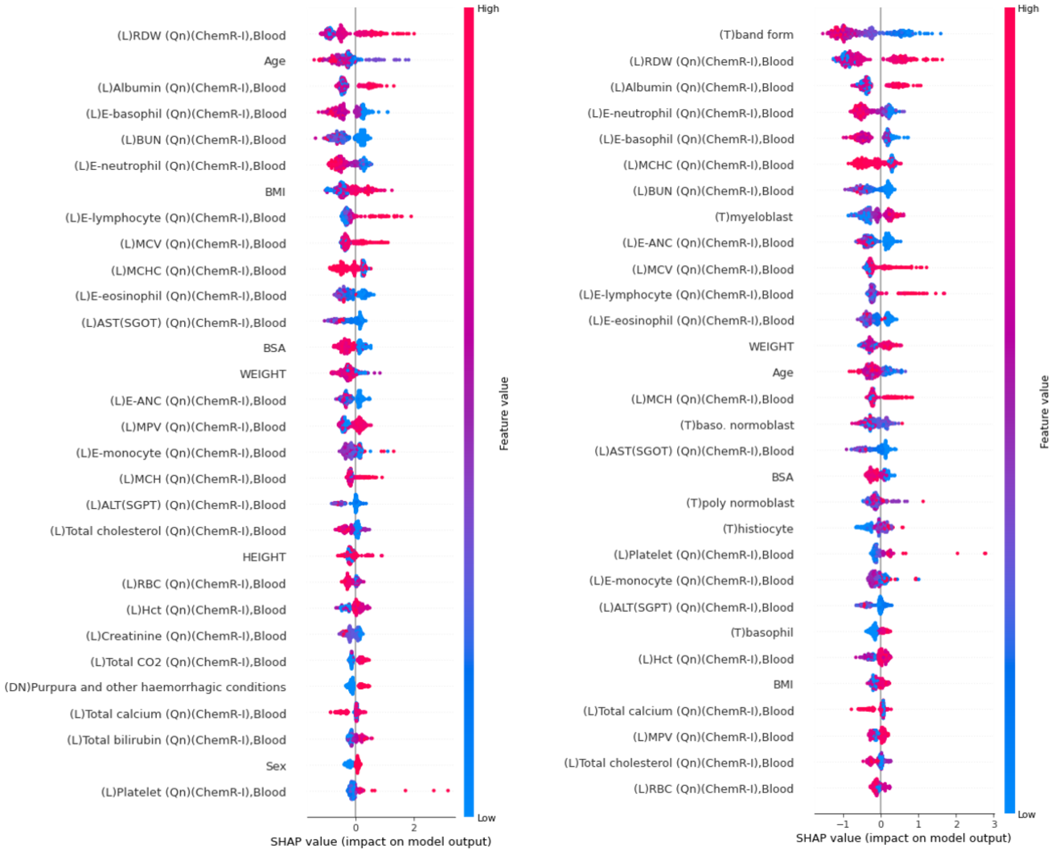
Figure 5. Receiver operating characteristic curve of the machine learning models; FPR: false positive rate; LR: logistic regression; ROC: receiver operating characteristic; SVM: support vector machine; TPR: true positive rate; XGB: extreme gradient boosting

Explainable ML model

The Shapley value is widely utilized in cooperative game theory and serves as a means to achieve equitable allocation outcomes for model interpretability. Notably, it provides a valuable tool for assessing the importance of individual features and understanding their influence on model decisions. These attributes play a pivotal role in enhancing the interpretability of machine learning models [31]. In this study, we conducted an analysis using the XGBoost (XGB) model to identify the top 30 Shapley values in both the base dataset and the base+text dataset, as illustrated in Figure 6.

In the Figure 6 legend, '(L)' represents laboratory test results, '(DN)' represents diagnoses received at the reference time, and '(T)' represents numerical value data extracted from textual sources.

In the base dataset, which was exclusively trained on fundamental structured data, it became evident that test outcomes such as red blood cell distribution width (RDW) and E-lymphocyte results played a pivotal role in the model's predictive capacity. Additionally, physical information had an influence on predictions. However, in the model incorporating the base+text dataset, which includes textual data, it became apparent that features derived from text sources, such as band form and myeloblast, assumed an important role in model prediction. This underscores the wealth of significant clinical information contained within textual data, thereby contributing positively to model performance.



<base dataset> **<base+text dataset>**

Figure 6. The Shapley Value Results for Each XGB Model; BMI: body mass index; BSA: body surface area; RDW: red blood cell distribution width;

Discussion

In this research, we have constructed a predictive model for disease progression in patients diagnosed with Idiopathic Cytopenia of Undetermined Significance (ICUS) who were admitted to the Asan Medical Center in Seoul. The principal objective of this study was to leverage a wide array of Electronic Medical Record (EMR) data from these patients to discern and forecast the critical clinical factors that impact the progression of myeloid malignancies in ICUS.

Throughout this procedure, we integrated crucial clinical information found within the unstructured text data of EMR. This strategic inclusion had a favorable effect on enhancing the performance of our predictive model. Notably, the model's reliability for predicting ICUS progressing to myeloid malignancy was bolstered by effectively refining and amalgamating genetic information, a key variable, with the unstructured text data. Our findings substantiate that the model's predictive accuracy significantly benefits from the incorporation of text data.

Nonetheless, these investigations are subject to certain limitations. First, in the process of excluding individual patient cohorts, each diagnosis was exclusively defined based on diagnosis codes. Given the potential absence of diagnostic codes within the EMR, it is customary to incorporate medications or test results that may co-occur when assembling patient cohorts. However, due to the limited knowledge regarding specific characteristics or treatments that influence disease progression in ICUS, the dataset was primarily structured to encompass solely the diagnostic records of patients. To mitigate the potential of this approach leading to incomplete patient groups, patient cohorts were structured with guidance from expert chart reviews conducted by clinicians.

Second, this study encountered challenges related to an imbalance in target labels. Specifically, Label 1 represents patients with ICUS progressing to myeloid malignancy, making up only 2.5% of the overall dataset. This label imbalance is a factor that can contribute to overfitting when applying machine learning techniques. To address these issues, we employed Stratified 10-fold cross-validation technology to assess the model's consistent prediction capability across various data subsets, thereby demonstrating the stability and reliability of the model.

Third, the language model utilized in this study underwent fine-tuning using a subset of textual results from patients with specific medical conditions. Although we employed language models trained on life sciences and medical terminology, the limitations inherent in such data sources may raise concerns regarding the ability to capture a broader range of clinical language. In our future research endeavors, we aspire to enhance the development of a more comprehensive disease prediction model by exploring language models that incorporate diverse textual data sourced from EMR.

Conclusion

In conclusion, our study successfully established a prediction model for disease progression from ICUS to myeloid malignancy, demonstrating its potential utility as a clinical prediction tool. Additional integration of textual clinical information, such as genetic data, can improve model performance and provide clinical insights for ICUS patient management.

Conclusions

In the first study, we utilized the electronic medical records (EMR) from Asan Medical to investigate the clinical characteristics of 16,934 patients with underlying atherosclerotic cardiovascular disease (ASCVD) who subsequently underwent reevaluation of their low-density lipoprotein cholesterol (LDL-C) levels. Our objective was to offer real-world evidence concerning the influence of early LDL-C level reduction on major adverse cardiovascular events (MACE) recurrence rates and medical costs. Subsequently, we projected the MACE recurrence rate over a 20-year follow-up period for these patients, computed the ensuing medical costs, and conducted an analysis by categorizing them into groups based on their LDL-C levels. This enabled us to furnish tangible clinical evidence regarding the benefits of early LDL-C goal achievement.

In the second study, we introduced a machine learning model designed to predict disease progression in idiopathic cytopenia of undetermined significance (ICUS) patients and explored the clinical factors that influence it. We integrated standardized patient characteristics from EMR into the model. Additionally, a pre-trained language model (PLM) was fine-tuned using chromosomal text data and bone marrow test result reports to capture crucial clinical features embedded within the text, which were then incorporated into the model. The XGBoost model ultimately achieved an AUROC of 0.817 and employed Shapley values to visualize the key factors that impact disease progression. These models are anticipated to contribute to the future development of medical artificial intelligence models based on EMR and medical text data.

Finally, our study provided clinical evidence supporting the early reduction of LDL-C levels to prevent cardiovascular disease recurrence using EMR data. Additionally, we introduced a machine learning model for disease prediction in ICUS patients based on EMR text data. The utilization of EMR data in real-world research and the development of artificial intelligence models can contribute to the establishment of reliable treatment guidelines and support medical decision-making. Furthermore, by expanding this research, we anticipate enhancing the utility of clinical text information within EMR and developing a more comprehensive disease prediction model through the creation of a large-scale language model trained on various types of text data in EMR.

eTable 1. The definition of ASCVD (MI, Unstable Angina, Stable Angina, Asymptomatic CAD, IS/TIA, PAD)

Inclusion criteria	
Myocardial infarction	***Patients who meet all of the following conditions after January 1, 2000:*** **1. Patients with recorded ICD diagnosis codes.** **<ICD-10>** I21 (Acute myocardial infarction) I22 (Subsequent myocardial infarction) I23 (Certain current complications following acute myocardial infarction) **2. Hospitalized patients including admissions through the emergency room.** **3. Patients who underwent CAG or PCI within 7 days of diagnosis code registration date.**
Unstable Angina	***Patients who meet all of the following conditions after January 1, 2000:*** **1. Patients with recorded ICD diagnosis codes.** **<ICD-10>** I20.0 (Unstable Angina) **2. Hospitalized patients including admissions through the emergency room.** **3. Patients who underwent CAG or PCI within 7 days of diagnosis code registration date.**
Stable Angina	***Patients who meet all of the following conditions after January 1, 2000:*** **1. Patients with recorded ICD diagnosis codes.** **<ICD-10>** I20.8 (Other forms of angina pectoris) I20.9 (Angina pectoris, unspecified) I24 (Other acute ischaemic heart diseases) I25.2 (Old myocardial infarction) I25.3 (Aneurysm of heart) I25.4 (Coronary artery aneurysm and dissection) I25.5 (Ischaemic cardiomyopathy) I25.8 (Other forms of chronic ischaemic heart disease) I25.9 (Chronic ischaemic heart disease, unspecified)

	2. Hospitalized patients including admissions through the emergency room. **3. Patients who underwent CAG or PCI within 7 days of diagnosis code registration date.**
Asymptomatic coronary artery disease	***Patients who meet all of the following conditions after January 1, 2000:*** **1. Patients with recorded ICD diagnosis codes.** **<ICD-10>** I25.0 (Atherosclerotic cardiovascular disease, so described) I25.1 (Atherosclerotic heart disease) I25.6 (Silent myocardial ischaemia) **2. CAG test results indicating moderate or severe findings or CCTA test results with values of 50 or higher.** ***Patients with coronary artery stenosis of 50% or greater as determined by CAG or CCTA, who also meet other ASCVD conditions, will be excluded.***
Ischemic stroke	***Patients who meet all of the following conditions after January 1, 2000:*** **1. Patients with recorded ICD diagnosis codes.** **<ICD-10>** I63 (Cerebral infarction) G45.9 (Transient cerebral ischaemic attack, unspecified) **Excluded haemorrhagic stroke** I60 (Subarachnoid haemorrhage) I61 (Intracerebral haemorrhage) I62 (Other nontraumatic intracranial haemorrhage) **Excluded TIA** G45 (Transient cerebral ischemic attacks and related syndromes G46 (Vascular syndromes of brain in cerebrovascular diseases) **2. Patients with at least one brain CT or MR imaging results within 30 days of diagnosis code registration date.** **Excluded Lacunar infarction stroke**

	Excluding individuals with the terms ‘Lacunar’, ‘Small Vessel’, or ‘SVD’ mentioned in the stroke note considering stroke subtype classification information.
PAD	***Patients who meet all of the following conditions after January 1, 2000:*** **1. Patients with recorded ICD diagnosis codes.** **<ICD-10>** I70 (Atherosclerosis) I73 (Other peripheral vascular diseases) I74 (Arterial embolism and thrombosis) **2. Patients who have undergone surgical procedures for arterial thrombosis and thrombosis, peripheral vascular procedures, or vascular angiography within 7 days of diagnosis code registration date.**

eTable 2. The definition of the assessment of the study subject' baseline clinical conditions.

Condition	ICD-10 code
Chronic Kidney Disease	**Patients who satisfy either the A or B condition:** A. N18 (chronic kidney disease) B. eGFR equal to or lower than 90.
Diabetes mellitus	**Patients who satisfy either the A or B condition:** A. E10 ~ E14 (Diabetes mellitus) B. HbA1c equal to or greater than 6.5%.
Metabolic Syndrome	Individuals satisfying two or more of the following criteria will be included: 1. Triglycerides equal to or greater than 150. 2. For males: HDL levels less than 40; for females: HDL levels less than 50. 3. SBP equal to or greater than 130, or DBP equal to or greater than 80. 4. Glucose equal to or greater than 100.
Hypertension	**Patients who satisfy either the A or B condition:** A. I10 (Essential (primary) hypertension) I11 (Hypertensive heart disease) I12 (Hypertensive renal disease) I13 (Hypertensive heart and renal disease) I15 (Secondary hypertension) B. Individuals with a history of prescription for: 1. Beta-blocker 2. RAAS inhibitor 3. Calcium channel blocker
Congestive heart failure	I42 (Cardiomyopathy) I43 (Cardiomyopathy in diseases classified elsewhere) I50 (Heart failure)
Atrial fibrillation disease	I48 (Atrial fibrillation and flutter)
Cancer	C00 ~ C97 (Malignant neoplasms)
Inflammatory Disease	
Rheumatoid arthritis	M05 (Felty syndrome) M06 (Other rheumatoid arthritis)
Psoriasis	L40 (Psoriasis)

HIV	B20 ~ B24 (Human immunodeficiency virus [HIV] disease)

eTable 3. Medication codes of study subjects' baseline medication use.

Drug class	Active ingredient(s) in medication
Statin	atorvastatin, cerivastatin, fluvastatin, lovastatin, pitavastatin calcium, pravastatin sodium, rosuvastatin, simvastatin
Ezetimibe	ezetimibe
Fibrate	bezafibrate, fenofibrate, gemfibrozil
Niacin	acipimox
Cholestyramine	cholestyramine resin
PCSK9 inhibitor	aliroumab, evolocumab
Aspirin	aspirin
P2Y12 inhibitor	clopidogrel, prasugrel, ticagrelor, ticlopidine hcl,
Beta-blocker	arotinolol hcl, atenolol, bevantolol, bisoprolol fumarate, carvedilol, celiprolol hcl, nebivolol, propranolol, propranolol hcl, s-atenolol
RAAS inhibitor	alacepril, benazepril, candesartan cilexetil, captopril, cilazapril, enalapril, enalapril maleate, eprosartan mesylate, fimasartan, fimasartan potassium, fosinopril, imidapril, irbesartan, lisinopril, losartan, moexipril hydrochloride, olmesartan, olmesartan medoxomil, perindopril tert-butylamine, quinapril, ramipril, telmisartan, temocapril, valsartan, zofenopril, zofenopril calcium
Calcium channel blocker	amlodipine, benidipine, cilnidipine, diltiazem hcl, felodipine, lacidipine, nicardipine, nifedipine, nilvadipine, nimodipine, nisoldipine, nitrendipine, s-amlodipine

eTable 4. The definition of Outcome variables (All-cause mortality, MI, IS, Hospitalization for UA, CR)

Outcome variable	
All-cause mortality	***Patients who meet all of the following conditions after index date:*** **1. Patients with documented in-hospital death dates and cancer-related death dates recorded at the hospital.**
Cardiac enzyme (Criteria used for selecting the outcome variable, though not the main outcome)	***Patients who meet all of the following conditions after index date:*** **1. Patients who have had Troponin-I or CK-MB tests measured from the time of admission, including through the emergency room, up to before PCI procedure.** **2. Patients with Troponin-I or CK-MB test results surpassing the upper limit of the reference range.** **3. In case of multiple tests for CAG, PCI or CABG on the same date of visit, the earliest test result will be considered.**
Myocardial infarction	***Patients who meet all of the following conditions after the index date:*** **1. Patients with recorded ICD diagnosis codes.** **<ICD-10>** I21 (Acute myocardial infarction) I22 (Subsequent myocardial infarction) I23 (Certain current complications following acute myocardial infarction) **2. Patients with recorded notes for CAG, PCI or CABG procedures.** **3. Cardiac enzyme during hospitalization.** ***Cases of MI occurring within 4 weeks from the previous event date will be treated as censored.**
Ischemic stroke	***Patients who meet all of the following conditions after the index date:*** **1. Patients with recorded ICD diagnosis codes.** **<ICD-10>** I63 (Cerebral infarction) G45.9 (Transient cerebral ischaemic attack, unspecified) **2. Patients with a brain CT or MR imaging results within 30 days after the index date.** ***Cases of Stroke/TIA occurring within 4 weeks from the previous event date will be treated as censored.**
Hospitalization for unstable angina	***Patients who meet all of the following conditions after the index date:*** **1. Patients who satisfy either the A or B condition:** **A. During the tracking period for outcome occurrence, if the newly added diagnosis includes the following codes:**

	<ICD-10> I20.0 (Unstable angina) I24.0 (Coronary thrombosis not resulting in myocardial infarction) I24.9 (Acute ischemic heart disease, unspecified) **Excluded chronic stable angina** I25 (Chronic ischemic heart disease) **B. In patients with predefined MI, if the maximum measured troponin I value during the outcome tracking period is less than 1.5.** **2. Patients with a brain CT or MR imaging results within 30 days after the index date.** **3. Patients with recorded notes for CAG, PCI or CABG procedures.** **4. Cardiac enzyme during hospitalization.**
Coronary Revascularization (PCI or CABG)	***Patients who meet all of the following conditions after the index date:*** **1. Patients in whom PCI or CABG procedures.**

References

1. Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Alonso, A., Beaton, A. Z., Bittencourt, M. S., ... & American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. (2022). Heart disease and stroke statistics—2022 update: a report from the American Heart Association. Circulation, 145(8), e153-e639.

2. Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., ... & GBD-NHLBI-JACC Global Burden of Cardiovascular Diseases Writing Group. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. Journal of the American College of Cardiology, 76(25), 2982-3021.

3. Punekar, R. S., Fox, K. M., Richhariya, A., Fisher, M. D., Cziraky, M., Gandra, S. R., & Toth, P. P. (2015). Burden of first and recurrent cardiovascular events among patients with hyperlipidemia. Clinical cardiology, 38(8), 483-491.

4. Baigent, C., Blackwell, L., Emberson, J., Holland, L. E., Reith, C., Bhala, N., ... & Collins, R. (2010). Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomized trials. The Lancet, 376(9753), 1670-1681.

5. Silverman, M. G., Ference, B. A., Im, K., Wiviott, S. D., Giugliano, R. P., Grundy, S. M., ... & Sabatine, M. S. (2016). Association between lowering LDL-C and cardiovascular risk reduction among different therapeutic interventions: a systematic review and meta-analysis. Jama, 316(12), 1289-1297.

6. Jellinger, P. S., Handelsman, Y., Rosenblit, P. D., Bloomgarden, Z. T., Fonseca, V. A., Garber, A. J., ... & Zangeneh, F. (2017). American Association of Clinical Endocrinologists and American College of Endocrinology guidelines for management of dyslipidemia and prevention of cardiovascular disease. Endocrine Practice, 23, 1-87.

7. Damman, P., van't Hof, A. W., Ten Berg, J. M., Jukema, J. W., Appelman, Y., Liem, A. H., & de Winter, R. J. (2017). 2015 ESC guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: comments from the Dutch ACS working group. Netherlands Heart Journal, 25, 181-185.

8. Grundy, S. M., Stone, N. J., Bailey, A. L., Beam, C., Birtcher, K. K., Blumenthal, R. S., ... & Yeboah, J. (2019). 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation, 139(25), e1082-e1143.

9. Miura, T., Izawa, A., Motoki, H., Miyashita, Y., Kashima, Y., Ebisawa, S., ... & Ikeda, U. (2015). Clinical impact of rapid reduction of low-density lipoprotein cholesterol level on long-term outcome of acute myocardial infarction in the statin era: subanalysis of the ALPS-AMI study. PLoS One, 10(6), e0127835.

10. Suvarna, V. R. (2018). Real world evidence (RWE)-Are we (RWE) ready?. Perspectives in clinical research, 9(2), 61.

11. Schubert, J., Lindahl, B., Melhus, H., Renlund, H., Leosdottir, M., Yari, A., ... & Hagström, E. (2021). Low-density lipoprotein cholesterol reduction and statin intensity in myocardial infarction patients and major adverse outcomes: a Swedish nationwide cohort study. European heart journal, 42(3), 243-252.

12. Lee, J., Lee, S. H., Kim, H., Lee, S. H., Cho, J. H., Lee, H., ... & Kim, J. H. (2021). Low-density lipoprotein cholesterol reduction and target achievement after switching from statin monotherapy to statin/ezetimibe combination therapy: Real-world evidence. Journal of Clinical Pharmacy and Therapeutics, 46(1), 134-142.

13. Kim, J., Park, K. T., Jang, M. J., Park, T. K., Lee, J. M., Yang, J. H., ... & Hahn, J. Y. (2018). High-intensity versus non-high-intensity statins in patients achieving low-density lipoprotein cholesterol goal after percutaneous coronary intervention. Journal of the American Heart Association, 7(21), e009517.

14. Shin, S. Y., Lyu, Y., Shin, Y., Choi, H. J., Park, J., Kim, W. S., & Lee, J. H. (2013). Lessons learned from development of de-identification system for biomedical research in a Korean Tertiary Hospital. Healthcare Informatics Research, 19(2), 102-109.

15. Valent, P., Horny, H. P., Bennett, J. M., Fonatsch, C., Germing, U., Greenberg, P., ... & Wells, D. A. (2007). Definitions and standards in the diagnosis and treatment of the myelodysplastic syndromes: Consensus statements and report from a working conference. Leukemia research, 31(6), 727-736.

16. Bennett, J. M. (2005, August). A comparative review of classification systems in myelodysplastic syndromes (MDS). In Seminars in oncology (Vol. 32, pp. 3-10). WB Saunders.

17. Ridgway, J. P., Lee, A., Devlin, S., Kerman, J., & Mayampurath, A. (2021). Machine learning and clinical informatics for improving HIV care continuum outcomes. Current HIV/AIDS Reports, 18, 229-236.

18. Segura-Bedmar, I., Colon-Ruiz, C., Tejedor-Alonso, M. Á., & Moro-Moro, M. (2018). Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. Journal of biomedical informatics, 87, 50-59.

19. Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data processing and text mining technologies on electronic medical records: a review. Journal of healthcare engineering, 2018.

20. Liao, K. P., Cai, T., Savova, G. K., Murphy, S. N., Karlson, E. W., Ananthakrishnan, A. N., ... & Kohane, I. (2015). Development of phenotype algorithms using electronic medical records and incorporating natural language processing. bmj, 350.

21. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

22. Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., ... & Salimi-Khorshidi, G. (2020). BEHRT: transformer for electronic health records. Scientific reports, 10(1), 7155.

23. Greenberg, P., Cox, C., LeBeau, M. M., Fenaux, P., Morel, P., Sanz, G., ... & Bennett, J. (1997). International scoring system for evaluating prognosis in myelodysplastic syndromes. *Blood, The Journal of the American Society of Hematology*, *89*(6), 2079-2088.

24. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

25. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine, 4(1), 86.

26. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), 1-23.

27. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

28. Durgesh, K. S., & Lekha, B. (2010). Data classification using support vector machine. Journal of theoretical and applied information technology, 12(1), 1-7.

29. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

30. Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

31. Rozemberczki, B., Watson, L., Bayer, P., Yang, H. T., Kiss, O., Nilsson, S., & Sarkar, R. (2022). The shapley value in machine learning. arXiv preprint arXiv:

국문 요약

전자 의무 기록을 활용한 임상 연구는 다양한 형태의 의료 데이터를 포함하며, 최근에는 자연어 처리 기술의 발전으로 인해 전자 의무 기록에 포함된 텍스트 데이터를 활용하는 연구도 활발하게 진행되고 있다. 이러한 연구는 실제 환경에서의 실사용 증거(Real-World Evidence, RWE)를 제공하며, 인공지능 기술과 결합해 질병 예측, 임상 의사 결정 지원 등 다양한 분야에 기여할 수 있다.

이에 따라 다음과 같은 목적의 연구를 계획하였다. 첫째, 심혈관 질환 입원 환자의 전자 의무 기록을 활용하여 고위험 심혈관 환자군에서 LDL-C 수치의 조기 목표 달성 여부가 심혈관 질환 재발률과 의료 비용에 미치는 영향에 대한 실사용 증거를 제시하는 것을 목적으로 한다. 둘째, 전자 의무 기록 내의 텍스트 데이터에 포함된 의료 정보를 포함하여 임상적중요성불명 특발혈구감소증(ICUS)의 질병 진행 가능성과 위험 요인을 예측하는 것을 목적으로 한다.

첫 번째 챕터에서, 심혈관 질환 환자의 전자 의무 기록을 추출하고 조기 LDL-C 수치 목표 달성에 따라 두 가지 그룹으로 나누어 분석하였다. 임상적 근거에 따라 투약 이력, 동반질환 등의 환자 특성을 정의하고 MACE 재발을 분석하였으며, 이로 인해 발생하는 의료 비용의 차이에 대한 연구를 수행하였다. 결과적으로 LDL-C 조기 감소를 달성한 그룹의 심혈관 재발 위험률이 낮게 나타났으며, 이는 조사된 의료 비용에서도 동일하였다. 이후 통계적 검증을 통해 연구 결과를 설명하였다.

두번째 챕터에서, ICUS 환자의 전자 의무 기록을 활용해 질병 예측 머신 러닝 모델을 개발하였다. 자연어 처리 기법을 통해 중요한 임상적 특성이 포함된 유전체 텍스트 데이터와 골수 검사 결과지를 전처리 하였으며, 질병 진행 예측을 위해 10개의 교차 검증을 사용하여 3개의 모델을 비교하였다. 이를 통해 예측 모델 개발과 위험 요인 분석을 수행하고 시각화 하였다. 최종적으로 텍스트 임베딩 데이터를 반영한 XGB 모델이 0.817로 가장 높은 AUROC 점수를 기록하였으며, Shapely value를 활용해 질환 예측에 반영된 주요한 위험 요인을 분석하였다.

결과적으로, 두 가지 연구를 통해 실제 환경에서의 질환 예후와 의료 비용을 분석하고, 구조화되지 않은 임상 텍스트 데이터를 활용하여 질병 진행을 예측하는 기계 학습 모델을 개발하였다. 전자 의무 기록 내 다양한 형태의 임상 정보를 활용한 연구로 확장함으로써, 환자 치료 가이드라인을 지원하여 임상 의사 결정과 의료 서비스 품질 향상에 기여할 것으로 기대된다.