공학석사 학위논문

Real-World Evidence (RWE) from EMR data and Development of Medical Artificial Intelligence Models

EMR 데이터를 활용한 RWE 임상연구수행 및 의료 인공지능 모델의 개발과 활용

울산대학교 대학원
의과학과
김민경

Real-World Evidence (RWE) from EMR data and Development of Medical Artificial Intelligence Models

지도교수 김영학

이 논문을 공학석사학위 논문으로 제출함

2024년 02월

울산대학교 대학원

의과학과

김민경

김민경의 공학석사학위 논문을 인준함

심사위원 이계화 (인)

심사위원 김영학 (인) 김영학인

심사위원 전태준 (인) 전태준인

울산대학교 대학원

2024 년 02 월

Abstract

Background

With recent advancements in healthcare-related technology, there has been a notable increase in the accumulation of electronic medical records (EMR) data across various medical institutions. Real-world evidence (RWE) research leveraging anonymized EMR data plays a crucial role in utilizing actual patient data to identify fundamental factors, relationships, and predictive risk factors. Of particular significance, cardiovascular disease (CVD) stands as one of the primary global causes of mortality, with a high lipoprotein(a) fraction being a major contributor to the heightened risk of CVD-related events.

Moreover, while the One-Hot Encoding (OHE) method is commonly employed for processing EMR data, it's worth noting that EMR data is predominantly recorded in the form of unstructured text data. Extracting valuable information from this textual data has become increasingly vital. Recent advancements in traditional Natural Language Processing (NLP) technology and word embedding methods have proven highly valuable, addressing the limitations of previous research by considering patients' treatment methodologies.

Indeed, this research is anticipated to harness data analytics techniques applied to electronic medical record (EMR) data, thereby paving the way for novel opportunities in healthcare innovation, with the potential to impact disease prevention and patient care significantly.

Objectives

The primary objective of this study is twofold. Firstly, it aims to harness EMRs for the purpose of conducting RWE research, grounded in authentic patient data. Specifically, the primary focus is to elucidate the connection between Lp(a) levels and cardiovascular outcomes within a cohort of high-risk CVD patients, while concurrently forecasting patient risk factors. Consequently, we intend to conduct a comprehensive study that estimates clinical characteristics and cardiovascular outcomes in correlation with Lp(a) levels, employing EMR data from individuals with a history of high-risk atherosclerotic cardiovascular disease (ASCVD) in Korea.

The second goal is aimed at improving the performance of medical artificial intelligence (AI) models. To achieve this, the study seeks to develop and validate a code embedding methodology that effectively captures patients' diagnostic patterns utilizing EMR data. The ultimate aim is to bolster the performance of medical artificial intelligence models.

Methods

Chapter 1. Epidemiology of lipoprotein(a) and the risk of MACE in ASCVD patients in South Korea:

This study was conducted on adult patients with ASCVD who visited Asan Medical Center in Seoul, South Korea, and underwent Lp(a) testing from January 1, 2001, to December 31, 2020. The collected data from ABLE were anonymized, and the structured data included information such as patients' basic information (age, height, weight, BMI), blood pressure measurements, admission and discharge dates, and visit type. Unstructured data included a variety of information, including surgical details, test results or interpretation, smoking status, and medication-related information. For patient-related variables (age, gender, BMI, smoking status), data from the same or closest date before the first Lp(a) measurement were used, and laboratory test results (blood pressure, cholesterol level, etc.) were collected 1 year from the date of index creation. Values measured within one year were selected.

The primary endpoint consisted of myocardial infarction, ischemic stroke, and all-cause mortality, while secondary endpoints included additional factors such as hospitalization for unstable angina.

Chapter 2. Cognizant Embeddings of ICD Codes via BERT: Leveraging Patient Diagnostic Patterns from a Large-scale Cardiovascular EMR Repository:

In this study, we employed fine-tuning on the pre-trained BERT MLM model using diagnostic codes and proposed a methodology to enhance the model's learning performance by reducing the dimensionality of the codes. Code embedding technology is a preprocessing technique that involves training an artificial intelligence model by embedding diagnostic codes, effectively converting words into numerical vectors.

This data encompassed patient records from Asan Medical Center in Seoul between January 1, 2000, and December 31, 2019. This yielded a total of 1,052,890 patients, including 572,811 from the CardioNet DB and 480,040 newly extracted patients. The extracted data included visit and discharge records, medication, diagnosis codes, and diagnosis dates.

A preprocessing step involved converting each patient's diagnosis code into a single code sequence to capture the visit unit effectively. Subsequently, partial sequences were pre-generated from the data to align with the BERT MLM framework.

We systematically explored variations in model dimensions and embedding pooling strategies to evaluate the model's efficiency. To evaluate the effectiveness of our code embedding method, we assessed the impact of different code subsequence alignment methods on the BERT MLM model's performance. In addition, we employed the xgb model to predict subsequent heart disease, allowing us to directly compare the performance of the OHE method with the Code Embedding method. Additionally, we used the t-SNE algorithm to visualize whether the model utilizing the code embedding method effectively captured the relationships between diagnosis codes.

Results

Chapter 1. Epidemiology of lipoprotein(a) and the risk of MACE in ASCVD patients in South Korea:

The study analyzed data from a final study population of 27,686 individuals who underwent Lp(a) testing between 2000 and 2020. These participants were divided into quintiles (Q1, Q2, Q3, Q4, and Q5) based on their Lp(a) levels. The high Lp(a) group (Q5) tended to be older, had a history of ASCVD (excluding stable angina), and showed higher total cholesterol and LDL-C levels. In addition, the highest Lp(a) group (Q5) was associated with older age, a history of ASCVD, higher total cholesterol, and LDL-C levels. The patients were categorized into five quintiles based on Lp(a) levels, and significant differences were observed between the highest and lowest quintiles. The 10-year cumulative incidence of MACE was approximately 29.5% in the entire cohort with a history of ASCVD. Kaplan-Meier curves demonstrated that the absolute risk of MACE recurrence increased with higher Lp(a) levels over a 10-year follow-up period.

Chapter 2. Cognizant Embeddings of ICD Codes via BERT: Leveraging Patient Diagnostic Patterns from a Large-scale Cardiovascular EMR Repository:

The proposed frequency-based sorting method outperformed alternative sorting approaches by reducing the loss by more than 0.1. Furthermore, a code embedding model trained in a 128-

dimensional space exhibited outstanding predictive performance in forecasting I50 (congestive heart failure), achieving an impressive AUROC value of approximately 0.997. In contrast, the XGB model using the OHE approach yielded a significantly lower AUROC value of 0.840, indicating suboptimal performance for this specific task. We observed that despite a substantial reduction in dimensionality, the 'Code Embedded XGB Models' achieved an AUC of 0.96, which is approximately 0.1 higher than the 'OHE XGB Models'. Particularly in real-world clinical predictions of MACE within one year for patients who underwent PCI or CABG, our embedding method reduced dimensions by about 96.5% compared to OHE and demonstrated an approximately 6% improvement in disease prediction performance.

Additionally, t-SNE visualizations confirmed that related diagnostic codes were located in similar two-dimensional vector spaces, revealing a tendency for diseases related to the same clinical groups to cluster closely based on classification results.

Conclusions

This study highlights the crucial role of adapting to the growing medical data volume and emerging healthcare technologies, benefiting disease prediction and treatment for real patients.

First, this study involved the construction of a patient cohort based on Lp(a) levels using EMR data from Korean ASCVD patients. It employed various statistical methods to validate the correlation between Lp(a) levels and the occurrence of recurrent MACE, providing valuable insights into critical clinical characteristics for individuals diagnosed with ASCVD.

Second, a methodology to enhance medical AI model performance was proposed by utilizing EMR data and developing a code embedding model. These findings suggest that our code sequence alignment method better understands important patient information in EMRs through NLP-based context embedding and strengthens models that identify associations with clinical diseases such as diagnostic codes or medications.

Finally, this study underscores opportunities for disease prevention and treatment through the application of RWE research utilizing EMRs. In addition, this research has the versatility to be applied to various clinical studies by integrating other unstructured EMRs. These results emphasize that medical code-embedded management, which integrates multiple data sources, is applicable to various medical prediction models and underscores the potential for risk prediction using realistic real-world datasets. Future research endeavors are anticipated to broaden and enhance medical AI models built upon EMRs, paving the way for the development of more advanced pretrained language models that leverage extensive EMR text data.

Keywords: electronic medical records, lipoprotein(a), ASCVD, cardiovascular diseases, MACE, ICD-10, diagnosis, natural language processing, code embeddings, BERT, one-hot encoding, representation learning.

Contents

Abbreviation

AI: artificial Intelligence
ABLE: Asan Biomedical research environment
AMC: Asan medical center
ASCVD: atherosclerotic cardiovascular disease
AUROC: area under the receiver operating characteristic
BMI: body mass index
CABG: coronary artery bypass grafting
CAD: coronary artery disease
CHD: coronary heart disease
CKD: chronic kidney diseases
CNN: convolutional neural network
CVD: cardiovascular disease
DB: data base
DM: Diabetes mellitus
DL: deep learning
eGFR: estimated glomerular filtration rate
ELMo: embeddings from language model
EHR: electronic health record
EMR: electronic medical record
GBM: gradient boosting algorithm
GPT: generative pre-trained transformer
HF: heart failure
HTN: hypertension
ICD: international classification of diseases of world health organization
ICU: intensive care unit
INDT: date of visitation or admission
INNO: patient encounter number
IS: ischemic stroke
KM: Kaplan-Meier
LDL-C: low-density lipoprotein cholesterol
LDL-C$_{\text{corr:}}$ low-density lipoprotein cholesterol corrected for lipoprotein(a)
LOS: length of stay
Lp(a): lipoprotein(a)
MACE: major adverse cardiovascular events
MI: myocardial infarction
MIMIC: medical information mart for intensive care
ML: machine learning
MLM: masked language model
MRI: Magnetic resonance imaging
NLP: natural language processing
OHE: one-hot encoding
OUDT: date of discharge
PAD: peripheral artery disease
PAID: patient identification
PCI: percutaneous coronary intervention

PLM: pre-training language model
PTCA: percutaneous transluminal coronary angioplasty
RCS: restricted cubic spline
RFECV: recursive feature elimination with cross-validation
ROC: receiver operating characteristic
RNN: recurrent neural network
TIA: transient ischemic attack
t-SNE: t-distributed stochastic neighbor embedding
W2V: word2vec
XGB: extreme gradient boosting

List of Tables

List of Figures

Introduction

Due to recent advances in healthcare-related technology, the amount of clinical data collected by medical institutions is rapidly increasing. Electronic medical records (EMR) are a type of clinical data and contain various medical records of patients. These data are de-identified through a pseudonymization or anonymization process to protect personal information and are used to develop artificial intelligence models in the medical field.

In addition, it plays an important role in RWE (Real-World Evidence) research that understands the complex mechanisms of diseases and develops treatments for diseases that were difficult to treat with existing methods. RWE research contributes to clearly understanding the relationship between conditions and clinical factors such as diseases, drugs, and side effects based on actual patient data and predicting patient risk. Hence, retrospective studies using large-scale EMR data can perform various risk prediction studies and be used as RWE in the medical field.

In particular, cardiovascular disease (CVD) is recognized as one of the leading causes of death worldwide, and high Lipoprotein(a) levels are known to be one of the major factors that increase the risk of CVD-related events. Cardiovascular disease is one of the acute and chronic diseases accompanied by various comorbidities and requires continuous and active management. Therefore, we would like to conduct a study to estimate clinical characteristics and cardiovascular outcomes according to Lp(a) levels using EMR in patients with a history of high-risk ASCVD in Korea.

Moreover, in the context of these clinical studies, it is customary to handle numerical data in Electronic Medical Records (EMR) using the prevalent One-Hot Encoding (OHE) method. Nevertheless, a significant portion of EMR data is captured in the form of unstructured text, which poses limitations when extracting valuable insights from medical texts. Recently, there has been a surge of interest in the application of Natural Language Processing (NLP) techniques, particularly word embedding, to process text data and incorporate it into artificial intelligence models. However, it is crucial to note that existing research has been hampered by a lack of consideration for patients' diagnosis patterns. To address this limitation, we expect to develop a model utilizing the diagnosis code embedding method. This innovative approach effectively captures and represents the intricacies of patients' diagnosis patterns, ultimately enhancing the performance of medical artificial intelligence models.

Hence, our research encompasses two distinct studies: the first focusing on Real-World Evidence (RWE) clinical research using the aforementioned EMR data, and the second centered around the development and utilization of a medical artificial intelligence model.

Firstly, the primary objective of the initial chapter is to conduct Real-World Evidence (RWE) studies employing authentic hospital Electronic Medical Record (EMR) data. This involved predicting Lp(a) levels and the likelihood of recurrent Major Adverse Cardiovascular Events (MACE) in patients hospitalized for cardiovascular diseases. EMR data were meticulously sourced from the Asan Medical Center's (AMC) ABLE Database for patients with Atherosclerotic Cardiovascular Disease (ASCVD). The extracted clinical data pertaining to cardiovascular diseases encompassed both outpatient and hospitalization records, along with specialized test results like echocardiograms and stroke records.

Additionally, during the data extraction process, we conducted a meticulous data-cleansing procedure, systematically identifying and eliminating outliers and erroneous entries that could potentially compromise the integrity of the dataset. Furthermore, we facilitated natural language processing technology to incorporate unstructured data sources, particularly test result records, into our statistical risk factor verification process. A study population of 29,868 people was formed according to Lp(a) quintiles, and statistically significant clinical variables were investigated through in-depth clinical discussions. We also utilized a variety of advanced statistical techniques, including survival analysis and Receiver Operating Characteristic (ROC) curve analysis. As a result, it was confirmed that the higher the Lp(a) level in ASCVD patients, the more significant the correlation with the occurrence of recurrent MACE.

Secondly, the primary objective of the second chapter is to enhance the efficacy of medical artificial intelligence models through the creation of a diagnostic code embedding method that adeptly captures the diagnostic patterns of patients. To leverage genuine patient diagnosis information and disease diagnosis data standardized by ICD-10 codes, various datasets including patient hospitalization and discharge records, diagnosis codes, diagnosis names, and diagnosis dates were systematically extracted from the Electronic Medical Records (EMR) sourced from the Asan Medical Center (AMC) ABLE Database. Subsequently, a code sequence, thoughtfully designed to encapsulate the chronological aspects of these data elements, was meticulously crafted.

For the dataset comprising generated code sequences, a model was crafted to optimize the embedding dimension utilizing a BERT-based PLM model. Subsequently, the XGB model was employed to predict 10 specific diagnosis codes pertaining to heart diseases, and its performance was scrutinized. It was affirmed that the model employing the code embedding method outperformed the OHE encoding XGB model in predicting subsequent heart-related diseases. Furthermore, employing dimensionality reduction visualization techniques, we ascertained that the diagnosis code embedding model effectively captures intricate relationships among crucial diagnosis codes in medical data.

This study is anticipated to contribute significantly to the understanding of disease causes and mechanisms, as well as the prediction of risk factors for real patients through Real-World Evidence (RWE) clinical research using Electronic Medical Records (EMR). Furthermore, it establishes the groundwork for enhancing the efficiency of future EMR-based research by employing low-dimensional compression techniques and fine-tuning diagnostic codes with the pre-trained BERT model. This effort lays the foundation for improving the overall performance of medical artificial intelligence models based on EMR data. It holds substantial potential to positively impact and enhance the effectiveness of future clinical research.

Building upon these achievements, our future aim is to develop a medical artificial intelligence model using a pre-trained language model, drawing from a wide range of EMR text data sources. This endeavor is expected to create new opportunities in the fields of medical research and patient care.

Chapter 1. Epidemiology of lipoprotein(a) and the risk of MACE in ASCVD patients in South Korea

Introduction

background

Cardiovascular disease (CVD) stands as a leading global cause of mortality, and its incidence has been steadily rising in recent times [1][2]. Numerous studies have indicated that individuals who have experienced a cardiovascular event face an elevated risk of additional events. Therefore, reducing risk factors associated with CVD is essential in mitigating the occurrence of cardiovascular complications [3]. Traditionally recognized risk factors for CVD include age, gender, smoking status, systolic blood pressure, total cholesterol, and levels of low/high-density lipoprotein cholesterol.

However, recent research has expanded our understanding of CVD risk factors to include neutral fat, homocysteine, fibrinogen, and lipoprotein cholesterol. Additionally, factors associated with inflammatory responses, such as Lipoprotein(a), have gained attention as emerging risk factors for coronary artery disease [4]. Of particular interest, Lipoprotein(a) levels have been identified as an independent risk factor for atherosclerosis, and there is growing focus on their role in elevating the risk of MACE (Major Adverse Cardiovascular Events), particularly as Lp(a) levels increase.

Lipoprotein(a) is a plasma lipoprotein characterized by its unique structure, composed of LDL particles containing cholesterol ester, neutral fat, apolipoprotein B-100 (apoB-100), and apolipoprotein(a) (apo(a)) [5]. Elevated levels of Lp(a) play a role as an anti-fibrinogen by inhibiting the binding of plasminogen and fibrin. Plasminogen and fibrin share structural similarities, but only plasminogen possesses fibrinolytic activity. When this inhibition occurs, it can lead to an increased risk of cardiovascular (CV) complications by promoting arteriosclerosis through thrombogenic and antifibrinolytic effects [6]. Notably, when an increase in Lp(a) levels coincides with a rise in low-density lipoprotein cholesterol concentration, it further escalates the risk of coronary artery disease and other vascular diseases [7].

In international settings, academic research is consistently being conducted to investigate the distribution of Lp(a) across various demographics, including age, gender, and race, and its associations with other risk factors for cardiovascular disease [8][9][10]. Furthermore, the European Atherosclerosis Society (EAS) has put forth the proposition that Lp(a) levels may elevate the risk of cardiovascular disease (CVD) development, even when traditional risk factors have been optimized [11]. It's important to note that while an individual's Lp(a) level tends to remain relatively stable over time, there is notable heterogeneity in Lp(a) levels among different population groups, including variations across various racial backgrounds. For instance, in certain African American populations, the median Lp(a) levels can be up to four times higher compared to white populations [12][13].

The consideration of East Asian/Korean populations in a multi-ethnic study suggests adjusting population-specific Lp(a) risk for CHD [14]. While there is a correlation between Lp(a) and CHD risk in Chinese subjects, a subsequent Korean cardiac catheterization (PC)-based study revealed that patients with Lp(a) levels > 30 mg/dL were originally categorized as Lp(a) ≤ 30 mg/dL. It was confirmed that the risk of MACE increased by 1.17 times compared to the previously reported findings [15].

Presently, Lipoprotein(a) is recognized as a primary cardiovascular disease (CVD) risk factor, and extensive research is being carried out through various prospective randomized controlled trials (RCT) [16][17][18]. These studies have substantiated that interventions aimed at regulating Lp(a) levels or employing LDL apheresis techniques hold the potential to mitigate CVD risk[19]. However, RCT face limitations related to participant diversity, as they tend to select specific matches that may not fully represent real-world clinical scenarios. Consequently, there is a pressing need for cohort studies, particularly those encompassing minority populations, to comprehensively evaluate risk factors and bridge the existing gaps in CVD incidence [20].

Objectives

The primary objective of this study is to investigate the epidemiology of lipoprotein(a) (Lp(a)) and its association with major cardiovascular events (MACE) risk in the Korean population. Upcoming international outcome trials will evaluate the effectiveness of treatments aimed at reducing Lp(a) levels, but this study aims to bridge the gap between real-world settings (RWE) and clinical trial results. We sought to achieve this by examining cardiovascular risk in real-world patients with a history of cardiovascular disease (ASCVD) who met the inclusion criteria for the outcome trial.

Furthermore, the secondary purpose of this study is to perform Real-World Evidence (RWE) clinical research using authentic patient data extracted from electronic medical records (EMR).

Specifically, the primary goal is to outline the distribution of Lipoprotein(a) levels and the clinical characteristics of patients who underwent evaluation within one year before the initial Lp(a) screening date. For the secondary objective, a cohort study design was employed. We conducted a longitudinal cohort study to explore the relationship between Lp(a) levels and the risk of Major Adverse Cardiovascular Events (MACE).

The primary aim of this research is to establish the association between Lp(a) levels and cardiovascular outcomes and to predict risk factors in patients with a high risk of cardiovascular disease (CVD). This endeavor is expected to enhance our comprehension of the underlying mechanisms of the disease and the identification of risk factors in real patients. These endeavors are anticipated to address the limitations of randomized controlled trial (RCT) studies and enhance our understanding of results in real clinical settings. This study is expected to yield valuable insights.

Methods

Study Design

We tested the hypothesis that higher Lp(a) will be associated with a higher risk of MACE compared to those with lower Lp(a). For this purpose, we used a retrospective analysis of the South Korea population to describe the Lp(a) distributions and patient profiling, and to investigate the epidemiology of Lp(a) and the MACE risk. We identify the risk factors for elevated Lp(a) and describe Lp(a) levels and MACE outcomes in ASCVD population.

For the primary objective, a cross-sectional analysis was employed to depict the distribution of Lp(a) in the South Korean population and to assess patients' clinical characteristics within 1 year prior to the date of their initial Lp(a) screening.

For the Secondary objective, we used a cohort study design, conducted a retrospective analysis of a routinely collected EHR database of a tertiary hospital, Asan Medical Center (AMC). A longitudinal cohort study was to investigate the exposure-response relationship between Lp(a) level and the risk of MACE. The overall study design is described in the schema on Figure 1.

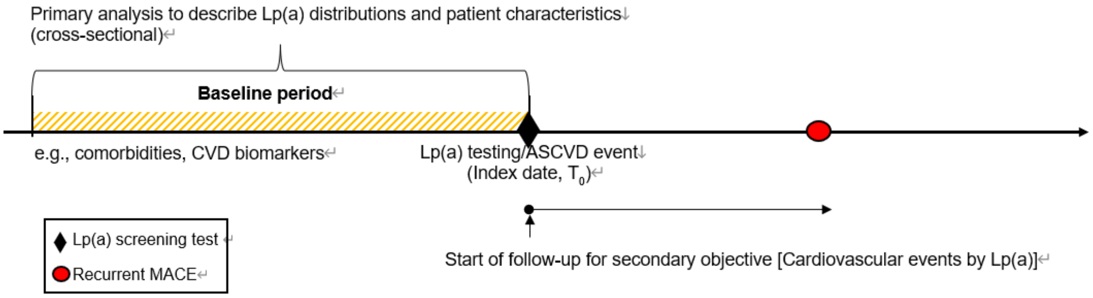

Primary analysis to describe Lp(a) distributions and patient characteristics
(cross-sectional)
Baseline period
e.g., comorbidities, CVD biomarkers
Lp(a) testing/ASCVD event
(Index date, T_0)
◆ Lp(a) screening test
● Recurrent MACE
Start of follow-up for secondary objective [Cardiovascular events by Lp(a)]

Figure 1. The study design schema of overall.

Study Population

This study was approved by the regional ethics committee in Asan Medical Center, Korea (IRB no. 2022-1173). The IRB conducted a risk/benefit analysis and determined that the study constitutes minimal-risk research. Thus, the IRB approved the waiver of the requirement to obtain informed consent.

This study was based on a real multicenter patient cohort. It included adult patients aged 18 years or older who were admitted to Asan Medical Center in Seoul, South Korea, with a previous history of ASCVD (Atherosclerotic Cardiovascular Disease) and who underwent Lp(a) testing between January 1, 2001, and December 31, 2020. The eligibility assessment and baseline period covered the one-year period preceding the index date (the initial Lp(a) screening).

Data description

Data acquisition

Asan Medical Center (AMC) is one of the largest tertiary hospitals in Korea with a large ASCVD research population and has data suitable for studying Lp(a)-related outcomes in Asia. ABLE is AMC's de-identified EHR database, which contains de-identified information on 4 million patients and is updated every three days. They also have a wealth of clinical information, including patient demographics (age, gender, date of death), lifestyle variables (smoking status, BMI), comorbidities (such as diabetes mellitus, chronic kidney disease, previous ASCVD subtypes, and cancer), as well as information on lipid-lowering treatment history, including statins, and laboratory results, which encompass LDL-C, estimated glomerular filtration rate (eGFR), blood pressure, levels of coronary artery stenosis, and C-reactive protein levels.

In order to monitor the continuous follow-up period of patients, EMR data was collected until December 31, 2021. Furthermore, to encompass the diagnostic, prescription, and surgical records of patients admitted through the emergency room, data extraction was comprehensive and covered single admissions, outpatient visits, as well as emergency room records.

We extracted structured and unstructured data, and all data were anonymized. The structured data include personal information such as age, height, weight, body mass index (BMI), blood pressure, measurement times, and dates. Additionally, visit data, including visit dates, admission and discharge dates, and visit types, were included. Diagnosis dates, diagnostic information, surgical or treatment dates, medication prescription dates, prescription codes, procedure codes, test dates, and test codes were also part of the structured data. The unstructured data comprise information on types of surgeries and tests, test and surgery result reports or interpretations, smoking status, active ingredients and dosages of medications, and more.

Data preprocessing

To document the hospitalization history of patients during the study period, we took into account each patient's individual hospitalization episodes, including admissions through the emergency room. The electronic medical record (EMR) in the ABLE database processes data by treating admissions to the emergency department as distinct events. However, to provide a comprehensive description of the ED admission, we utilized the admission and discharge timestamps recorded during the ED visit and processed it into one single admission record.

We utilized patient diagnostic information and mapped major diagnoses received during hospital visits to ICD-10 codes. We set a 7-day period centered on the patient's hospitalization date and used procedure records, including surgery and interventional procedures, to determine the history of cardiovascular revascularization in ASCVD patients. The level of stenosis was determined by extracting values from the test results records of coronary angiography or coronary CT angiography for the three major coronary arteries: left anterior descending artery (LAD), circumflex artery (LCX), and right coronary artery (RCA). Given that the test result record is presented as free text sentences without a structured format, we employed regular expressions, which are specific rules for extracting values related to test items such as LM, LAD, LCX, RCA, and RI. Afterwards, if the value of each item exceeded 50, it was classified as moderate or severe stenosis. We employed textual data extracted from the CT or MR imaging reports of patients who underwent these imaging modalities to identify individuals diagnosed with ischemic stroke. In order to meticulously differentiate and exclude cases of lacunar infarct strokes, we conducted exclusions based on the presence of specific terminology such as 'Lacunar,' 'Small Vessel,' or 'SVD' within the stroke records.

Using the constituents of prescribed medications identified in the EMR data, medication history was established. Specifically, for one of the major medications, statin preparations, they were categorized based on the duration of prescription. In cases where the prescription duration was equal, preference was given to high-intensity statin preparations over moderate and low-intensity ones. Additionally, when multiple medications were prescribed on the same day, prioritization was determined based on the total prescription days to reflect the patient's primary medication. The measurement of statin intensity involved the use of regular expressions to process text data, considering the components and dosage of the prescribed medication.

The definition of ‘cardiac enzymes during hospitalization’ was based on numerical data from patients with troponin-I or CK-MB results exceeding the upper limits. The upper limit thresholds for Troponin-I or CK-MB were calculated differently, considering the prescription practices at AMC. To ensure the consideration of only those cardiac enzyme measurements influenced during hospitalization, a standardization of time units was implemented according to the specified format and units was applied to all values, converting them to 'ng/mL'. The analysis exclusively incorporated test results obtained from the time of admission to the emergency room or hospitalization, up to just before the PCI procedure. In cases where multiple PCIs occurred within the same hospitalization period, the determination of the test date relied on either the closest date to the Troponin-I or CK-MB testing or the earliest date based on the hospitalization admission date. Furthermore, if multiple tests were conducted on the same day, the highest value of Troponin-I or CK-MB was chosen.

Key Variables

To investigate the association between Lp(a) levels and the incidence of Major Adverse Cardiovascular Events (MACE), Atherosclerotic Cardiovascular Disease (ASCVD) events were initially categorized into five distinct groups: Myocardial Infarction (MI) or Unstable Angina, Stable Angina, Atherosclerotic Coronary Artery Disease (ACAD), Ischemic Stroke/Transient Ischemic Attack (IS/TIA), and Peripheral Artery Disease (PAD). The electronic medical record (EMR) data used for ASCVD classification encompassed the period from January 1, 2000, to the index date. The definitions of ASCVD categories and their corresponding ICD-10 codes are as follows:

- Myocardial Infarction (MI) (ICD-10 codes, I21-I23) or Unstable Angina (ICD-10 code, I20) included patients admitted to hospitals, including those admitted through the emergency room. Patients who underwent Coronary Angiography (CAG) or Percutaneous Coronary Intervention (PCI) within 7 days from the date of diagnosis code registration were included.
- Stable Angina (ICD-10 codes, I20, I24-I25) encompassed patients admitted to hospitals, including those admitted through the emergency room. Patients who underwent Coronary Angiography (CAG) or Percutaneous Coronary Intervention (PCI) within 7 days from the date of diagnosis code registration were also included.
- Asymptomatic Coronary Artery Disease (ICD-10 code, I25) is defined as patients with moderate or severe Coronary Angiography (CAG) test results or those with ≥50% coronary

artery stenosis determined by Computed Tomography Coronary Angiography (CCTA) test results. Patients who met other ASCVD conditions were excluded.

- Ischemic Stroke (ICD-10 codes, I63, G45) comprises patients who had undergone at least one brain CT or MR imaging test within 30 days from the date of diagnosis code registration. Additionally, patients with the 'Lacunar' subtype mentioned in the stroke note were included, while individuals with the terms 'Small Vessel' and 'SVD' were excluded.
- Symptomatic Peripheral Arterial Disease (ICD-10 codes, I70, I73-I74) includes patients who underwent various surgical interventions related to arterial thrombosis, thrombosis surgery, peripheral vascular procedures, and angiography within 7 days from the date of diagnosis code registration.

In cases where multiple ASCVD events were diagnosed simultaneously on the same day, the patient was first categorized into the MI or Unstable Angina group, followed by Stable Angina and ACAD as appropriate.

To explore the association between Lp(a) levels and MACE incidence, we established the index date as the date of the first Lp(a) screening conducted after each patient's ASCVD diagnosis. The Lp(a) screening data encompassed the period from January 1, 2000, to December 31, 2020. This index date signifies the initial measurement point for Lp(a) when an ASCVD event was documented. Notably, Lp(a) numerical values were retained in their original form without any alterations or substitutions, in accordance with clinical deliberation.

For patient-related variables, including age, sex, BMI, smoking status, and others, we utilized data from the same date as the first Lp(a) measurement or the nearest available date before it. Smoking status was determined using nursing information survey records from the electronic medical record (EMR) database, allowing us to categorize specific smoking statuses. In the case of BMI, values falling below the 25th percentile and exceeding the 75th percentile were excluded from the analysis.

Laboratory test results, including SBP (Systolic Blood Pressure), DBP (Diastolic Blood Pressure), LDL-C (Low-Density Lipoprotein Cholesterol), total cholesterol, HDL-C (High-Density Lipoprotein Cholesterol), triglycerides, eGFR (estimated Glomerular Filtration Rate), and ACR (Albumin-to-Creatinine Ratio), were selected based on measurements taken within one year before or on the same date as the index date. If multiple test results were available for the same day, the maximum value was chosen. Outliers were excluded to focus on values falling within the clinically relevant range for each test parameter. In cases where LDL cholesterol values were missing, they were calculated using the Friedewald formula. Total cholesterol levels ranged from 20 to 600, triglycerides from 10 to 3000, HDL-C from 1 to 400, eGFR from 1 to 200, and ACR from 20 to 1000.

To account for the patient's concurrent medication information, the study considered medication histories up to 1 year before the index date. The accuracy of drug data was confirmed through clinical discussions, relying on AMC prescription codes and drug ingredient names. Comprehensive information about the ingredients of each medication can be found in eTable 1. This included the use of the following medications: statins, ezetimibe, other lipid-lowering agents, aspirin, p2y12 inhibitors, beta-blockers, RAAS (Renin-Angiotensin-Aldosterone System) inhibitors, estrogens, and calcium channel blockers.

Significantly, the study also took into account the intensity of statin use, especially in the context of assessing the impact of the Lp(a) fraction on the occurrence of MACE in patients with a history of ASCVD, where statin use is considered one of the risk factors. Specifically, for statin users, the study examined cases where there was documented history of statin use within the year prior to the index date to comprehensively capture the patient's medication history. To accomplish this, both the dates of drug usage and dosage information were meticulously considered.

Variables related to comorbidities encompassed cases diagnosed at least once based on the patient's medical history preceding the index date. Definitions of these comorbidities primarily relied on the utilization of ICD-10 codes and other pertinent patient data. Detailed information regarding each comorbidity definition can be found in eTable 2.

- Chronic Kidney Disease (ICD-10 code, N18) was defined as having an eGFR ≤90 before the index date.
- Hypertension (ICD-10 codes, I10-I13, I15) included individuals with a history of prescriptions for beta-blockers, RAAS inhibitors, and patients with a history of using one or more calcium channel blockers.
- Diabetes mellitus (ICD-10 code, N18) was defined as having HbA1c ≥6.5%.
- Other comorbidities comprised Congestive heart failure (ICD-10 codes, I42-I42, I50), Atrial fibrillation disease (ICD-10 codes, I48), Cancer (ICD-10 codes, C00-C97), and Liver Disease (ICD-10 codes, B18-B19, K70-K77).
- The definition of Metabolic Syndrome (MS) is as follows: Individuals with MS meet at least two or more of the following criteria:
    1. Triglycerides are equal to or greater than 150.
    2. For males: HDL levels are less than 40; for females: HDL levels are less than 50.
    3. SBP is equal to or greater than 130, or DBP is equal to or greater than 80.
    4. Glucose is equal to or greater than 100.

The primary endpoint was a composite of MI, ischemic stroke, and all-cause mortality. The secondary endpoint included a composite of MI, ischemic stroke, all-cause mortality, hospitalization for unstable angina. Major Adverse Cardiovascular Events (MACE) were defined as increase in cardiac enzymes during hospitalization, and reports of CT or MR imaging results, and documentation of coronary angiography (CAG) or percutaneous coronary intervention (PCI) or coronary artery bypass grafting (CABG) as well as attributed to International Classification of Diseases, Tenth Revision (ICD-10) codes related to cardiovascular disease.

Cardiovascular disease (ICD-10 codes, I20-I25, I63, G45.9) was divided into CHD (ICD-10 codes, I20-I25) and stroke (ICD-10 codes, I63, G45.9) in accordance with the AHA guidelines in eTable 3.

Patients with stenosis exceeding 50% or with moderate to severe findings, as indicated by CAG or CCTA test results, were also included in the baseline characteristic analysis.

Statistical Analysis

We conducted patient tracking from the date of the initial Lp(a) assessment until the earliest of the following events: the occurrence of Major Adverse Cardiovascular Events (MACE), censoring events, or the conclusion of the study period on December 30, 2021. Censoring events encompassed instances of follow-up failure, defined as the patient's last clinic visit, or cases where MACE events transpired in other medical facilities but the patients returned to the outpatient departments of Asan Medical Center (AMC), as recorded in the ABLE database. Notably, recurrent Myocardial Infarction (MI) or Ischemic Stroke (IS) events transpiring within the initial 4 weeks post-discharge for ASCVD were not considered as independent events but were uniformly censored.

We calculated the proportion and cumulative incidence of MACE across subgroups defined by quintiles of the Lp(a) mass distribution. Incidence functions were calculated using the Kaplan-Meier method. Cumulative incidence of MACE for each quintile of Lp(a) is shown. Hazard ratios (95% CI) for the risk of MACE were estimated using the Cox proportional hazards model.

We also used a restricted cubic spline (RCS) approach to estimate the association between Lp(a) levels and MACE risk. We estimated the association between Lp(a) levels and risk of MACE, characterized the exposure-response relationship, assessed between Lp(a) exposure groups defined by Lp(a) thresholds, and compared the risk of MACE.

Covariates considered included age (continuous, years), gender (categorical, male, female), ASCVD subtype (categorical, 0, 1), body mass index (BMI, continuous), and ASCVD-related baseline comorbidities (hypertension).) was included. , diabetes, chronic kidney disease, categorical, 0, 1), smoking status (categorical, never, past and current smoker), LDL-C and HDL cholesterol (continuous), history of statin use (categorical, high, moderate, low) and other lipid-lowering treatments (categories 0, 1). Additionally, if missing information exceeded 10%, we imputed missing data for potential confounders using multivariate imputation by chained equations (MICE). Covariates were carefully defined via literature review and clinically meaningful variables with consultations with the investigators.

The primary endpoint was defined as the recurrence of a composite of MI (Myocardial Infarction) or Unstable Angina, ischemic stroke, or all-cause mortality. The secondary endpoint was also defined as the recurrence of a composite of MI or Unstable Angina, ischemic stroke, or all-cause mortality.

R version 4.2.3 statistical software package was used for analysis. We defined 2-sided P values of <.05 as statistically significant.

Results

Primary analysis results

To characterize the distribution of Lp(a) within the Korean population and identify patient profiles correlated with elevated Lp(a), a histogram was utilized to illustrate the distribution of Lp(a) levels. The majority of patients exhibited Lp(a) levels within the range of 0 to 30.

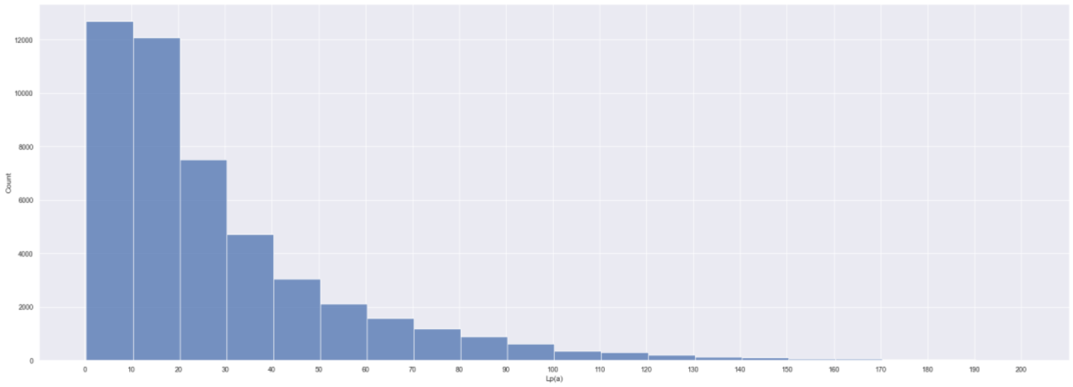


Figure 2. The histogram depicting Lp(a) distribution among patients who visited AMC between January 1, 2000, and December 31, 2020.

To characterize the distribution of Lp(a) within the Korean population and identify patient profiles correlated with elevated Lp(a) levels, we used histograms to illustrate the distribution of Lp(a) levels. As shown in Figure 2, the majority of patients had Lp(a) levels within the range of 0 to 30.

The eTable 4 displays the comparisons between quintile groups, taking into account age group (≥ 65 years), gender, LDL-C level, ASCVD status and subtype, preexisting comorbidities (e.g., DM, HTN, CKD), and smoking status (current smoker, former smoker, never smoker), as well as BMI category (underweight [<18.5 kg/m2], normal [18.5–22.9 kg/m2], overweight [23.0–24.9 kg/m2], obese I [≥30.0 kg/m2]). In comparison to Quintile 1, which had the lowest distribution of Lp(a) levels, Quintile 5 exhibited an average age approximately one year older (61.05 years versus 62.47 years) and LDL-C levels approximately 13 mg/dL higher (87.05 mg/dL versus 102.95 mg/dL). Moreover, there were more individuals diagnosed with Hypertension (6,932 versus 7,410), CHF (449 versus 540), and Rheumatoid arthritis (21 versus 30) in Q5 when compared to Q1.

Secondary analysis results

Between 2000 and 2020, a total of 47,818 patients aged 18 or older underwent Lp(a) screening at least once. From this group, 17,950 patients without a prior ASCVD diagnosis were excluded. Additionally, 182 patients with a diagnosis of hemorrhagic stroke and 273 patients with a diagnosis of lacunar infarction stroke prior to the index date were excluded. Furthermore, 1,727 patients with SBP > 180 or DBP > 110 at the time of Lp(a) screening were also excluded.

The final study population comprised 27,686 individuals, categorized into quintiles Q1, Q2, Q3, Q4, and Q5 based on Lp(a) levels. Specifically, the Q1 group included 5,449 individuals, Q2 had 5,566 individuals, Q3 consisted of 5,591 individuals, Q4 comprised 5,540 individuals, and Q5, with the highest distribution of Lp(a) levels, included 5,540 individuals. Additionally, each patient was further classified into five ASCVD subgroups (MI or Unstable Angina, Stable Angina, Asymptomatic Coronary Artery Disease, PAD), with some patients belonging to multiple subtypes. Figure 3 illustrates the selection process for the overall study cohort.

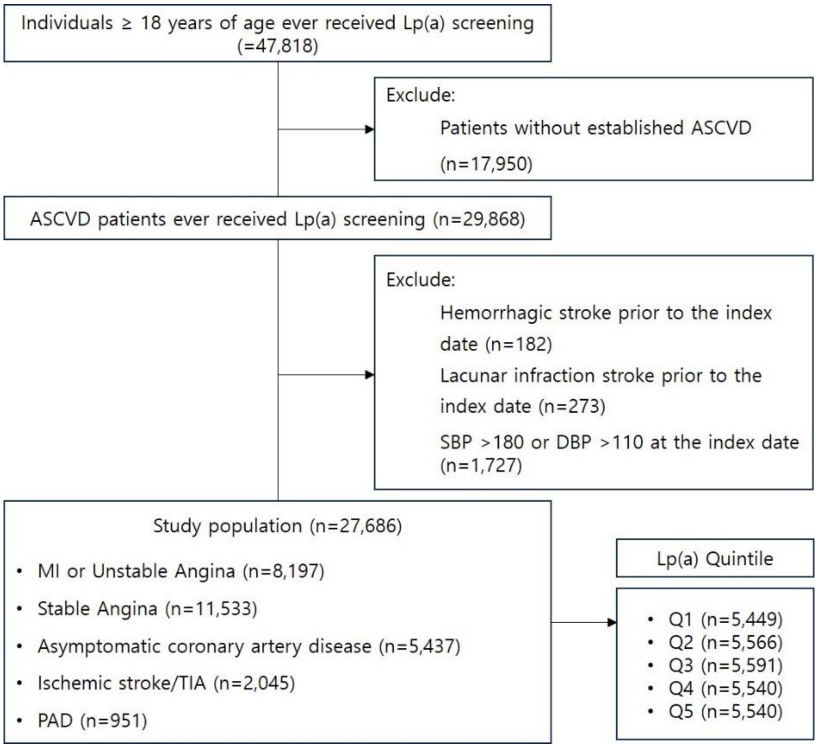

Individuals ≥ 18 years of age ever received Lp(a) screening (=47,818)
Exclude:
Patients without established ASCVD
(n=17,950)
ASCVD patients ever received Lp(a) screening (n=29,868)
Exclude:
Hemorrhagic stroke prior to the index date (n=182)
Lacunar infraction stroke prior to the index date (n=273)
SBP >180 or DBP >110 at the index date (n=1,727)
Study population (n=27,686)
• MI or Unstable Angina (n=8,197)
• Stable Angina (n=11,533)
• Asymptomatic coronary artery disease (n=5,437)
• Ischemic stroke/TIA (n=2,045)
• PAD (n=951)
Lp(a) Quintile
• Q1 (n=5,449)
• Q2 (n=5,566)
• Q3 (n=5,591)
• Q4 (n=5,540)
• Q5 (n=5,540)

Figure 3. Selection of the overall study cohort.

* Quintile ranges for Lp(a) were as follows: Q1 (0.4 ≤ Lp(a) < 8.3), Q2 (8.3 ≤ Lp(a) < 14.9), Q3 (14.9 ≤ Lp(a) < 24.9), Q4 (24.9 ≤ Lp(a) < 43.9), and Q5 (43.9 ≤ Lp(a) < 684).

In comparison to Quintile 1, which had the lowest distribution of Lp(a) levels, Quintile 5 consisted of older individuals who had a greater history of ASCVD (excluding stable angina) diagnoses. Furthermore, they had higher levels of total cholesterol (154 versus 167.7 mg/dL) and LDL-C (90.49 versus 101.87 mg/dL). Additionally, patients in Q5, compared to those in the Q1 group, were more likely to have comorbidities such as CKD (51.7% versus 51.2%), DM (30.3% versus 31.3%), HTN (81.3% versus 82.4%), and CHF (3.4% versus 4.4%). Notably, individuals taking statins after an ASCVD diagnosis had a high utilization rate of approximately 92% (including lipid-lowering treatments) in both Q1 and Q5.

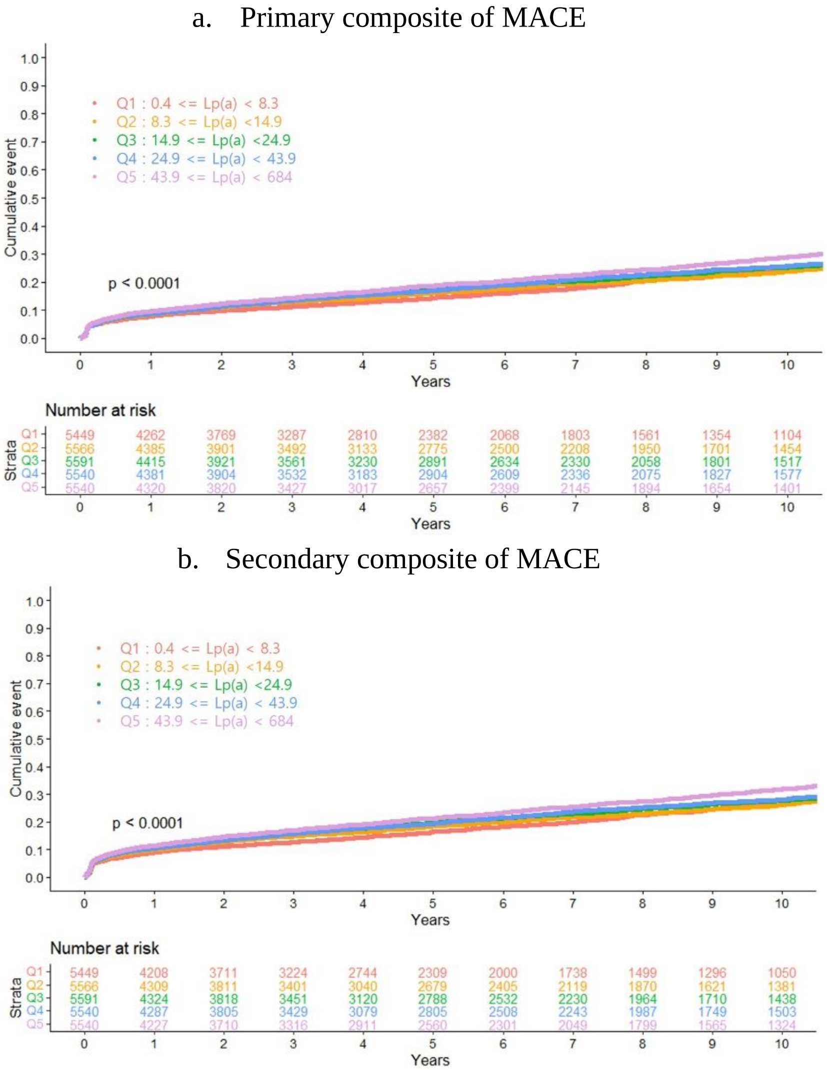
Table 1. Patient characteristics by Lp(a) quintiles in ASCVD patients.

	Quintile 1 (n:) [Lp(a) range: 0.4<=lpa<8.3]	Quintile 2 (n:) [Lp(a) range: 8.3<=lpa<14.9]	Quintile 3 (n:) [Lp(a) range: 14.9<=lpa<24.9]	Quintile 4 (n:) [Lp(a) range: 24.9<=lpa<43.9]	Quintile 5 (n:) [Lp(a) range: 43.9<=lpa<684]
Mean (SD) Age (years)	62.361 (10.71)	62.631 (10.63)	62.837 (10.42)	62.881 (10.32)	**63.395 (10.12)**
Age ≥65 y (%)	2377 (43.6)	2505 (45.0)	2603 (46.6)	2564 (46.3)	**2674 (48.3)**
Male	4037 (74.1)	3979 (71.5)	3914 (70.0)	3867 (69.8)	3628 (65.5)
History of ASCVD					
ASCVD subtype					
MI or Unstable Angina	1465 (26.9)	1737 (31.2)	1738 (31.1)	1657 (29.9)	**1600 (28.9)**
Stable Angina	2427 (44.5)	2280 (41.0)	2302 (41.2)	2285 (41.2)	2239 (40.4)
Asymptomatic CAD	1029 (18.9)	1065 (19.1)	1051 (18.8)	1102 (19.9)	**1190 (21.5)**
Ischemic stroke/TIA	428 (7.9)	368 (6.6)	407 (7.3)	405 (7.3)	**437 (7.9)**
PAD	174 (3.2)	190 (3.4)	188 (3.4)	188 (3.4)	**211 (3.8)**
Prior PCI/CABG	1706 (31.3)	1652 (29.7)	1625 (29.1)	1495 (27.0)	1419 (25.6)
Smoking status					
Never smoker	2812 (51.6)	2917 (52.4)	3025 (54.1)	2958 (53.4)	3137 (56.6)
Ex-smoker	1234 (22.6)	1204 (21.6)	1096 (19.6)	1137 (20.5)	1071 (19.3)
Current smoker	1146 (21.0)	1092 (19.6)	1100 (19.7)	1042 (18.8)	960 (17.3)
Mean (SD) BMI (kg/m^2)	24.996 (2.84)	24.9 (2.9)	24.8 (2.863)	24.7 (2.91)	24.564 (2.93)
Baseline Labs					
Mean (SD) SBP (mmHg)	125.112 (18.71)	124.113 (18.63)	124.3 (19)	124.411 (19.38)	125.134 (20.12)
Median (Q1-Q3) SBP (mmHg)	124 (112-137)	112 (110-136)	123 (110-136)	123 (110-136)	123 (110-138)

Mean (SD) DBP (mmHg)	73.297 (11.7)	73 (11.5)	72.874 (11.77)	73 (12.12)	72.9 (11.94)
Median (Q1-Q3) DBP (mmHg)	72 (65-80)	72 (65-80)	72 (65-80)	72 (65-80)	72 (65-80)
eGFR < 60 mL/min/1.73 ㎡	152 (2.8)	118 (2.1)	95 (1.7)	83 (1.5)	114 (2.1)
ACR ≥30mg/g	167 (3.1)	124 (2.2)	154 (2.8)	141 (2.5)	156 (2.8)
Mean (SD) Total cholesterol (mg/dL)	154 (39.84)	160.55 (39.42)	162.36 (40.61)	165 (41.63)	**167.7 (41.97)**
Median (Q1-Q3) Total cholesterol (mg/dL)	151 (124-180)	158 (132-187)	160 (134-188)	163 (135-191)	164 (137-193)
Mean (SD) LDL-C (mg/dL)	90.49 (34.10)	95.78 (34.21))	97.52 (35.15)	100.31 (37.04)	**101.87 (35.69)**
Median (Q1-Q3) LDL-C (mg/dL)	93(65-112)	93 (70-118)	94 (72-119)	96 (72-123)	**97 (75-124)**
Mean (SD) HDL-C (mg/dL)	43.22 (12)	42.68 (11.65)	42.5 (11.73)	42.68 (11.66)	43 (12)
Median (Q1-Q3) HDL-C (mg/dL)	42 (35-50)	41 (35-79)	41 (34-49)	41 (35-49)	42 (35-50)
Mean (SD) Triglycerides (mg/dL)	143.8 (95.28)	139.44 (83.21)	132.76 (72.38)	127.81 (72.29)	126.92 (66.1)
Median (Q1-Q3) Triglycerides (mg/dL)	121 (86-172)	119 (88-167)	115 (85-161)	113 (84-154)	112 (84-153)
Mean (SD) Lp(a) (mg/dL)	5.71 (1.47)	11.38 (1.9)	19.44 (2.87)	32.98 (5.4)	73.5 (30)
Median (Q1-Q3) Lp(a) (mg/dL)	24.92(23.14-26.78)	11.3 (9.7-13)	19.3 (17-21.8)	32.3 (28.2-37.3)	65.6 (52.6-85.125)
Level of stenosis					
N (%), CAG - Moderate stenosis (50-69%)	13 (0.2)	8 (0.1)	9 (0.2)	7 (0.1)	5 (0.1)
N (%), CAG - Severe stenosis (≥70%)	60 (1.1)	62 (1.1)	50 (0.9)	48 (0.9)	65 (0.4)

N (%), CCTA – Moderate or Severe	10 (0.2)	7 (0.1)	7 (0.1)	5 (0.1)	11 (0.2)
Chronic Kidney Disease	2818 (51.7)	2682 (48.2)	2662 (47.6)	2650 (47.8)	**2891 (52.2)**
Diabetes mellitus	1633 (30.3)	1627 (29.2)	1666 (29.8)	1584 (28.6)	**1732 (31.3)**
Hypertension	4430 (81.3)	4528 (81.4)	4560 (81.6)	4497 (81.2)	**4566 (82.4)**
Metabolic syndrome	4892 (89.8)	4892 (87.9)	4886 (87.4)	4766 (86.0)	4786 (86.4)
Congestive heart failure	186 (3.4)	188 (3.4)	176 (3.1)	220 (4.0)	**241 (4.4)**
Atrial fibrillation	232 (4.3)	257 (4.6)	223 (4.0)	216 (3.9)	194 (3.5)
Cancer	394 (7.2)	345 (6.2)	382 (6.8)	376 (6.8)	363 (6.6)
Liver disease	212 (3.9)	156 (2.8)	150 (2.7)	126 (2.3)	107 (1.9)
Lipid lowering treatments before the index date					
Statin					
High intensity	327 (6.0)	292 (5.2)	287 (5.1)	250 (4.5)	320 (5.8)
Moderate intensity	2765 (50.7)	2490 (44.7)	2567 (45.9)	2509 (45.3)	2647 (47.8)
Low intensity	64 (1.2)	75 (1.3)	54 (1.0)	69 (1.2)	77 (1.4)
Ezetimibe	15 (0.3)	7 (0.1)	16 (0.3)	7 (0.1)	23 (0.4)
Other lipid lowering treatments (fibrate, niacin, cholestyramine)	79 (1.4)	69 (1.2)	64 (1.1)	51 (0.9)	54 (1.0)
Aspirin or P2Y12 inhibitor before the index date	325 (6.0)	240 (4.3)	211 (3.8)	149 (2.7)	189 (3.4)
Beta-blocker before the index date	1201 (22.0)	1406 (25.3)	1500 (26.8)	1558 (28.1)	1563 (28.2)
RAAS inhibitor (ACE inhibitor, ARB, or aldosterone antagonist) before the index date	542 (9.9)	572 (10.3)	555 (9.9)	569 (10.3)	647 (11.7)
Calcium Channel Blocker	243 (4.5)	247 (4.4)	254 (4.5)	281 (5.1)	318 (5.7)

Estrogen/Hormone replacement therapy before the index date	61 (1.1)	60 (1.1)	48 (0.9)	55 (1.0)	68 (1.2)
Lipid lowering treatments after the ASCVD discharge date					
Statin					
High intensity	**483 (8.9)**	432 (7.8)	433 (7.7)	382 (6.8)	**436 (7.9)**
Moderate intensity	**4179 (76.7)**	4216 (75.7)	4275 (76.5)	4207(75.9)	**4224 (76.2)**
Low intensity	**157 (2.9)**	176 (3.2)	161 (2.9)	212 (3.8)	**227 (4.1)**
Ezetimibe	43 (0.8)	37 (0.7)	61 (1.1)	56 (1.0)	89 (1.6)
Other lipid lowering treatments (fibrate, niacin, cholestyramine)	182 (3.3)	186 (3.3)	148 (2.6)	165 (3.0)	154 (2.8)
Calendar Year (Index date)					
2001-2005	907 (16.6)	1332 (23.9)	1410 (25.2)	1531 (27.6)	1502 (27.1)
2006-2010	1402 (25.7)	1837 (33.0)	1938 (34.7)	1976 (35.7)	1841 (33.2)
2011-2015	1353 (24.8)	1259 (22.6)	1238 (22.1)	1136 (20.5)	1086 (19.6)
2016-2020	1787 (32.8)	1138 (20.4)	1005 (18.0)	897 (16.2)	1111 (20.1)



a. Primary composite of MACE

b. Secondary composite of MACE

Figure 4A. Cumulative incidence of recurrent MACE among by Lp(a) quintiles.

Cumulative incidence rates of primary and secondary endpoints in patients with a history of ASCVD are shown in Figure 4A. In the entire cohort, the 10-year cumulative incidence of the primary composite endpoint (MI or Unstable Angina, ischemic stroke, all-cause mortality) was approximately 29.5%, and the secondary composite endpoint (MI, hospitalization for Unstable Angina, ischemic stroke, all-cause mortality) was approximately 30.1%.

Furthermore, the absolute 10-year risk of recurrent MACE exhibited an increase as lipoprotein(a) levels rose. Particularly, the cumulative incidence of both primary and secondary endpoints during the 10-year follow-up period, as compared to the initial cumulative incidence, was notably higher in patients belonging to Q5, characterized by elevated Lp(a) levels, than in the Q1 group with lower Lp(a) levels.

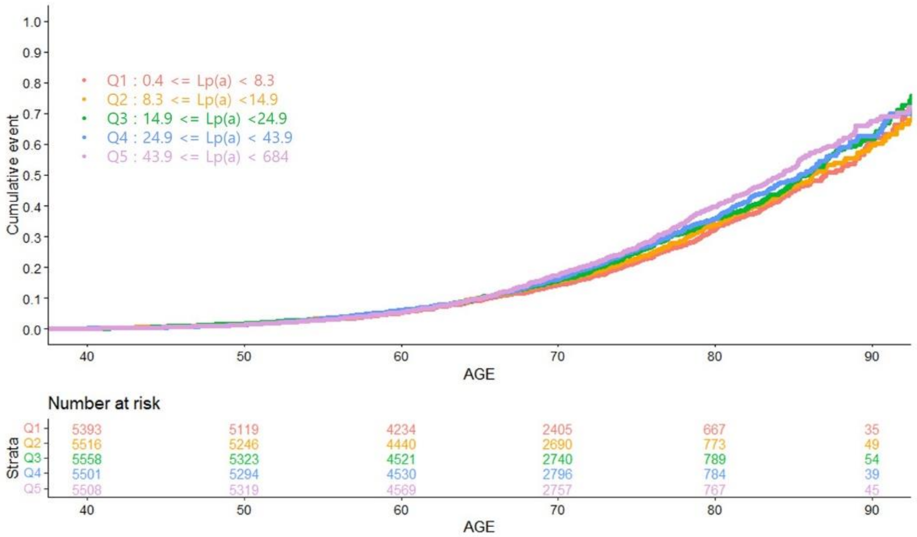


Figure 4B. Cumulative incidence of recurrent MACE (secondary composite MACE) by age among Lp(a) quintiles.

Since Lp(a) is generally expected to remain stable over time, we anticipate that baseline Lp(a) levels will not fluctuate during follow-up. Therefore, we analyzed the results by age group, starting from 40 years and older. The Kaplan-Meier (KM) results depicted in Figure 4B, which only included patients over 40 years of age with a history of ASCVD diagnosis, revealed that the absolute risk of MACE recurrence increased with age as lipoprotein(a) levels increased.

The cardiovascular event rates are presented in Table 2. When comparing the highest Lp(a) quintile group (Q5) and the lowest quintile group (Q1) among all patients who underwent Lp(a) screening and had a history of ASCVD, the 20-year incidence rate per 100 person-years for the primary composite endpoint was higher in patients with higher Lp(a) levels (3.85 versus 3.14). For the secondary composite endpoint, it was 4.41, which was approximately 0.9 higher than the rate of 3.50 in Q1. Among the individual components of the secondary endpoints, the 20-year incidence rates for the entire cohort were all higher in the Q5 group. Specifically, for all-cause mortality, Q1 was 1.01 and Q5 was 1.52; for MI, Q1 was 1.17 and Q5 was 1.29; for hospitalization due to unstable angina, Q1 was 0.47 and Q5 was 0.74; and for ischemic stroke (IS), Q1 was 1.03 and Q5 was 1.19.

Table 2. Association between Lp(a) level and outcomes comparing the highest Lp(a) quintile (Q5) with the lowest quintile (Q1).

Outcomes	Q1(n=5,449)			Q5(n=5,540)			Crude HR (95% CI)	Adjusted HR (95% CI)
	No of events	**Person years**	**Rate per 100 P-Y (95% CI)**	**No of events**	**Person years**	**Rate per 100 P-Y (95% CI)**		
Primary endpoint	940	29917.911	3.14	1293	33530.069	**3.85**	1.26 (1.16, 1.37)	1.19 (1.07, 1.33)
Secondary endpoint	1018	29081.100	3.50	1421	32190.669	**4.41**	1.3 (1.2, 1.41)	1.21 (1.1, 1.35)
Individual endpoints								
All-cause mortality	336	33218.039	1.01	579	38029.785	**1.52**	1.51 (1.32, 1.72)	1.42 (1.2, 1.68)
Myocardial infarction	363	30997.522	1.17	452	35022.255	**1.29**	1.15 (1, 1.32)	1.07 (0.89, 1.28)
Hospitalization for-unstable angina	150	31910.809	0.47	266	35832.832	**0.74**	1.65 (1.35, 2.01)	1.58 (1.17, 2.14)
Ischemic stroke	328	31760.555	1.03	428	35842.981	**1.19**	1.19 (1.03, 1.37)	1.15 (0.95, 1.38)

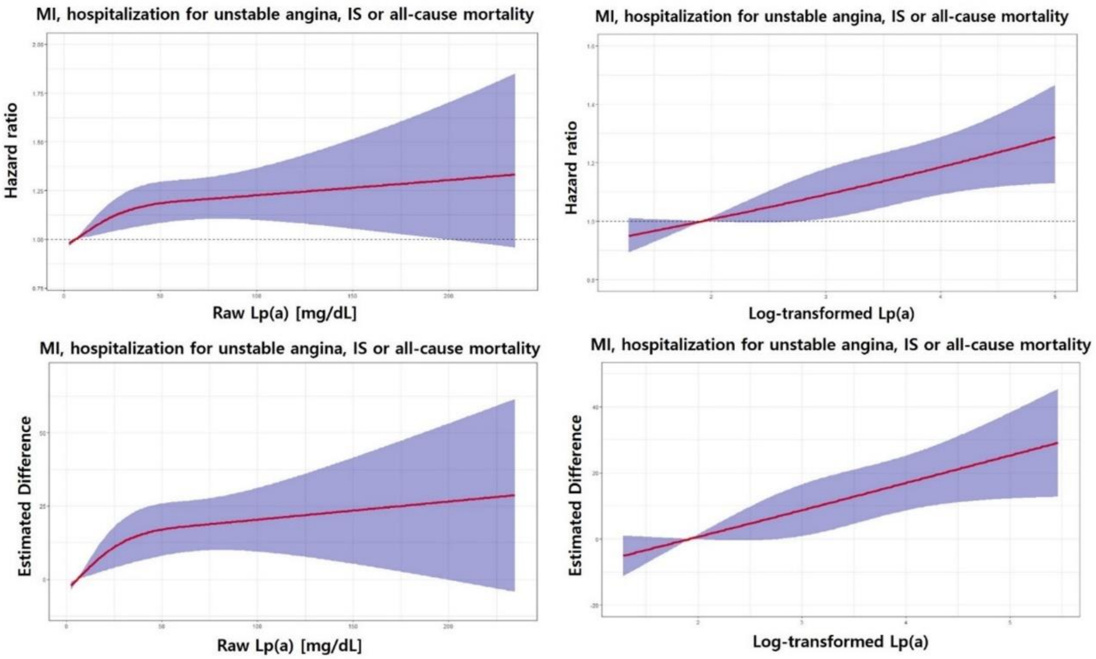


Figure 5. Restricted cubic spline analysis results exploring an exposure-response association between Lp(a) levels and the risk of MACE (Secondary endpoint).

Hazard ratios (HR; 95% CI) for the recurrence of MACE over a 20-year follow-up period, based on restricted cubic splines derived from a Cox proportional hazards model, were calculated while adjusting for age, sex, ASCVD subtype, body mass index, baseline comorbidities (hypertension, diabetes mellitus, chronic kidney disease), smoking status, LDL-C, HDL cholesterol, and the use of antiplatelet and statin/other lipid-lowering therapy at discharge.

Figure 5 illustrates the results, indicating that the risk of recurrent MACE was significantly lower in Q5 compared to Q1 after adjusting for risk factors. When considering the distribution of Lp(a) values in Figure 1, the risk ratio for recurrent MACE based on raw Lp(a) values exhibited a linear upward trend. Additionally, a graph reflecting the natural log-transformed Lp(a) values was presented. The results for log-transformed Lp(a) clearly demonstrated that the risk of MACE was significantly higher in the high Lp(a) level group (Q5) when adjusting for risk factors compared to the low Lp(a) level group.

Table 3. Subgroup analysis of the association between Lp(a) level and primary endpoint MACE (the highest Lp(a) quintile (Q5) versus the lowest quintile (Q1)).

Outcomes	**Q1**	**Q5**	**Crude HR (95% CI)**	**Adjusted HR* (95% CI)**
	No of events/No of subjects			
Age groups				
>65 years	475/2377	702/2674	1.18 (1.05, 1.33)	**1.2 (1.04, 1.38)**
Baseline LDL-C				
< 70 mg/dL	193/1368	125/735	1.21 (0.97, 1.52)	1.15 (0.91, 1.45)
≥ 70 mg/dL	466/3034	639/3130	1.26 (1.12, 1.42)	**1.24 (1.09, 1.39)**
Smoking				
Never	478/2812	717/3137	1.21 (1.08, 1.36)	1.24 (1.07, 1.43)
Ever	170/1234	213/1071	1.38 (1.13, 1.68)	1.34 (1.05, 1.71)
Current	259/1146	283/960	1.19 (1.01, 1.41)	**1.05 (0.85, 1.3)**
Diabetes mellitus				
Yes	332/1633	475/1732	1.3 (1.13, 1.5)	1.15 (0.97, 1.37)
No	608/3816	818/3808	1.24 (1.11, 1.37)	1.24 (1.09, 1.43)
Hypertension				
Yes	744/4430	1045/4566	1.28 (1.17, 1.47)	1.19 (1.05, 1.34)
No	196/1019	248/974	1.17 (0.97, 1.42)	1.24 (0.97, 1.59)
CKD				
Yes	494/2818	702/2891	1.27 (1.13, 1.42)	1.21 (1.06, 1.38)
No	446/2631	591/2649	1.23 (1.09, 1.39)	1.17 (0.97, 1.41)
Use of lipid-lowering treatment at baseline				
Yes	497/3186	629/3066	1.2 (1.07, 1.35)	1.18 (1.03, 1.36)
No	443/2263	664/2474	1.31 (1.16, 1.48)	1.21 (1.02, 1.44)

*Adjusted for age, sex, ASCVD subtype, body mass index, baseline comorbidities (hypertension, diabetes mellitus, chronic kidney disease), smoking status, LDL-C, HDL cholesterol, antiplatelet and statin/other lipid lowering therapy at discharge.

The analysis results of potential risk factors contributing to MACE outcomes based on Lp(a) levels within a 20-year observation period for the entire cohort are summarized in Table 3. During the total 20-year observation period, the risk of meeting the primary or secondary endpoint was significant in the group aged 65 or older (HR 1.2 (95% CI, 1.04-1.38)) and in the segment with LDL-C levels of 70 (HR 1.24 (95% CI, 1.09-1.39)) or higher. Regarding smoking status, it was observed that the adjusted hazard ratio (HR) was lower in patients who were currently smoking (HR 1.05 (95% CI, 0.85-1.3)), suggesting a potential broad interpretation depending on the timing of smoke status measurement, as there is a possibility that smokers may have ceased smoking after the initial ASCVD diagnosis.

Discussions

Our study has provided valuable insights into the patient profiles of individuals with elevated Lp(a) levels who are at heightened risk of future ASCVD events within the Asian population. The Asan Medical Center (AMC) possesses tailored data for outcomes research pertaining to Lp(a) in Asia, focusing on a substantial cohort of ASCVD patients who have undergone coronary angiography or percutaneous coronary intervention (PCI) since 2001, establishing it as one of the preeminent comprehensive medical centers in Korea systematically tracking Lp(a) measurements over a span of two decades. Hence, this information has the potential to serve as a crucial reference dataset for formulating strategic approaches in other Asian nations, such as China and Japan.

One noteworthy finding from our study is that among high-risk Korean ASCVD patients managed within routine clinical settings, those exhibiting elevated Lp(a) levels are significantly predisposed to recurrent cardiovascular events, as evidenced by a 20-year incidence rate (cumulative incidence rate) of 3.5 per 100 person-years. Generally, cardiovascular event rates in the United States research [21] were around 6.0 per 100 persons. According to Swedish research [22], the rate was 7.0 per 100 person-years over an average of 7.3 years of follow-up [15]. However, our study reports comparatively lower figures, and this discrepancy may be attributed, in part, to our selection of recurrent MACE occurrences as the study endpoint rather than initial incidents.

In this study, we did not find a significant relationship between the risk of meeting the primary or secondary endpoint criteria and statin intensity in the majority of patients who were using moderate-intensity statins. While previous research frequently demonstrates a discernible relationship between statin intensity and the risk of cardiovascular disease, it is worth noting that East Asian patients tend to exhibit a comparatively subdued response to statins in comparison to their Caucasian counterparts. This racial disparity likely contributes to the absence of a substantial relationship observed in this study.

Moreover, our observations indicate that advanced age is linked to an increased susceptibility to recurrent MACE events in individuals aged over 40, with Lp(a) levels as a contributing factor. Since Lp(a) levels are profoundly shaped by genetic factors and exhibit relative stability across diverse lifestyles, the prospects for extending our findings to broader research initiatives become more feasible.

This study has several strengths. It leveraged electronic medical record (EMR) data to conduct a large-scale real-world evidence (RWE) study and demonstrated that Lp(a) levels in the adult population in Korea with underlying ASCVD are indeed an important risk factor for recurrent MACE events. Consequently, the epidemiological and patient profiles obtained from this study will contribute significantly, particularly in understanding the results of the Phase III clinical trial of Olpasiran in the Asian population.

Limitations

This study has several limitations. First, there is a potential measurement error in Lp(a) values. AMC measures mass Lp(a) in mg/dL using Siemens immune nephelometric assays. However, the lack of assay standardization and sensitivity of most assays to Lp(a) size make comparisons between studies challenging, potentially leading to underestimation or overestimation of true Lp(a) levels [23][24]. Using isoform-insensitive assays for measuring Lp(a) molar concentration could reduce measurement errors, but such measurements were unavailable in the AMC ABLE dataset.

Second, risk factor analysis for factors that increase Lp(a) may be biased. While Lp(a) levels are largely genetically determined and stable, some non-genetic factors can influence them, potentially introducing bias into the risk factor analysis. Therefore, careful consideration of selection bias is necessary to avoid distortion of results.

Third, using all-cause mortality rates may introduce information bias. Current privacy regulations in Korea prevent access to statistics on causes of death. However, some research suggests that using all-cause mortality rates may be a better measure for assessing disease burden [25][26]. Therefore, using all-cause mortality rates is still considered a valid approach for measuring the burden of recurrent cardiovascular disease.

Fourth, individuals receiving Lp(a) tests at AMC are more likely to have cardiovascular-related symptoms or diseases. While it is not appropriate to generalize results from symptomatic patients to the general population, investigating the clinical characteristics of this large ASCVD study group is an important first step in understanding the epidemiology of Lp(a) in the East Asian population.

Fifth, this is an observational study, which has inherent limitations. The initial ASCVD events in the AMC system are used as index events, and all individuals are evaluated with a one-year assessment period. This may lead to missing information on the first event history if a patient experiences two events at three-year intervals but only the second event is recorded in the AMC database.

Finally, there are limitations in the utilization of ABLE data. Missing data in structured data were challenging to replace, and even with imputation methods, determining the attribution of patient characteristics was difficult. Additionally, most patient data in ABLE were in an unstructured free-text format, potentially requiring advanced Nature Language Processing (NLP) techniques. This might have led to an underutilization of the available unstructured data. Therefore, we plan to conduct future research utilizing EMR and incorporating numerous AI-based studies, including NLP techniques.

Conclusion

In Korean adults with a history of ASCVD, higher Lp(a) levels were associated with an increased risk of developing cardiovascular disease MACE. In adults aged 18 and above with a pre-existing ASCVD history, elevated Lp(a) levels may be linked to a higher risk of subsequent recurrent MACE.

Chapter 2. Cognizant Embeddings of ICD Codes via BERT: Leveraging Patient Diagnostic Patterns from a Large-scale Cardiovascular EMR Repository

Introduction

Background

In recent years, the number of patients in hospitals has rapidly increased due to advances in healthcare technology, resulting in a corresponding increase in various EMR [27] data. The EMRs contain vast repositories of patient medical information within hospital systems and play an important role in clinical and medical AI research [28][29]. As a result, there has been a growing focus on harnessing EMR data to enhance the performance of medical AI models [30]. Furthermore, the EMRs serve as a comprehensive source of information regarding patient demographics and past medical history, emphasizing the significance of extracting valuable insights from medical text.

The EMRs record and manage diagnostic and medications-related information, which are essential for documenting patient conditions in various unstructured code formats. Diagnostic codes play a crucial role in characterizing the patient conditions and other essential attributes. Furthermore, RWE studies often define target cohorts based on ICD-10th codes, depending on the study objectives. However, in large-scale EMR databases, patient information is often recorded in text format, making simple analysis difficult [31]. Therefore, developing methods to effectively extract medical information from coded EMR data could have a significant impact on the effectiveness of medical AI models.

Currently, EMRs such as ICD-10th codes, medication information, and treatments are primarily converted using the OHE method and integrated into deep learning (DL) models [32][33]. OHE is one of the most basic ways to represent words in vector form. However, the most significant constraint of OHE is that it cannot take into account the complexity of EMRs in text format. Using the OHE method on large-scale EMR data increases the diversity of codes, which can lead to the curse of dimensionality and data redundancy issues when training models. The primary concern is that these characteristics could potentially restrict the applicability of medical prediction models. This key limitation is that it fails to consider the similarities among the text data that make up the majority of the EMRs, resulting in the model ignoring the complex mechanisms of patient information. Therefore, there is a need for effective methods to extract medical text from EMRs.

Prior works

To overcome the limitations of OHE, word-embedding methods like the Skip-gram algorithm from Word2Vec (W2V) [34] have attracted considerable attention. W2V is a word-embedding method that represents words as continuous vectors and can calculate similarity between words. This algorithm is employed to organize ICD codes, test results, and medication codes within EMRs into a unified space, facilitating the creation of low-dimensional representations of medical concepts [35]. Specifically, the sliding window feature of Skip-gram can be utilized to clearly visualize code sequences, serving as a similarity matching method for patients within a cohort [36]. Word-embedding approaches also enable the association of diagnostic codes with medications, aiding in the identification of medications linked to specific diseases for treatment or prevention purposes. While W2V can handle text data that OHE cannot, it does not fully capture the complete context and unique characteristics of EMRs.

In recent years, considerable research has been conducted to enhance the performance of medical AI models by directly applying NLP mechanisms to overcome the challenges mentioned earlier [37]. This approach utilizes the context of surrounding words, causing the same word to be embedded differently depending on the context. For instance, consider sentences like 'The ocean's wave crashed against the shore with tremendous force.' and 'She used a simple hand wave to signal that everything was okay.' In these sentences, the word 'wave' assumes different meanings depending on the context. Contextual embedding assists in precisely determining the intended meaning of 'wave' by considering the surrounding words.

These improvements, especially considering context-sensitive embeddings, are often achieved by fine-tuning domain-specific data in pre-trained language models (PLM) such as BERT, T5, and GPT. Among these, the BERT pre-learning model created by pre-training external data such as MIMIC is often used in medical AI research [38][39]. However, studies utilizing MIMIC use only some of the frequently occurring codes, complicating rare disease prediction. In addition, models trained solely on MIMIC often overlook the actual patterns of patient diagnoses [40][41].

To increase the accuracy and efficiency of NLP-based medical AI models, it is essential to recognize and integrate real-world diagnostic patterns. This approach takes comprehensive disease complexity and important medical variables into account, enabling medical AI models to provide more accurate diagnoses, support comprehensive patient management, and informed treatment decisions. For example, the pre-training cohort of Med-BERT [42] consists of medical records of approximately 20 million real patients extracted from Cerner. The ability to utilize larger cohorts and extended visit sequences can help models understand more comprehensive context, and larger corpus sizes generally improve model performance. However, it must be taken into account that incorporating additional data into a PLM is a resource- and time-consuming task, as observed in previous studies [43]. Clinical prediction models must not only achieve high performance, but also be generally applicable in real clinical settings. In addition, since similar diagnostic codes are sometimes shared between patient groups diagnosed with different diseases, the need for a model that recognizes such diverse patient information to identify actual disease patterns and predict clinical problems is emphasized.

Objectives

Our aim is to construct a model that harnesses the diverse information available within EMRs to enhance the accuracy of patient diagnosis and treatment prediction. To achieve this goal, we propose an effective method for enhancing the performance of medical AI models through the utilization of EMR data. This research is anticipated to aid healthcare professionals in their decision-making processes by applying the generalization of medical AI models, ultimately contributing to improved patient treatment outcomes and healthcare practices.

Methods

Data description

Data acquisition

In this study, we utilized EMRs from patients who visited Asan Medical Center in Seoul (AMC) from January 1, 2000, to December 31, 2019. The EMR records from January 2000 to December 2016 were derived from the CardioNet DB [44], consisting of data from 572,811 individuals, with a specific focus on EMR data related to visits to the Departments of Cardiology and Thoracic Surgery at AMC. EMRs from January 1, 2017, to December 31, 2019, additionally include personal information on 189,413 individuals who visited AMC. Our comprehensive EMR database incorporates records from all other departments at AMC that patients visiting the Cardiology Department accessed during their visits.

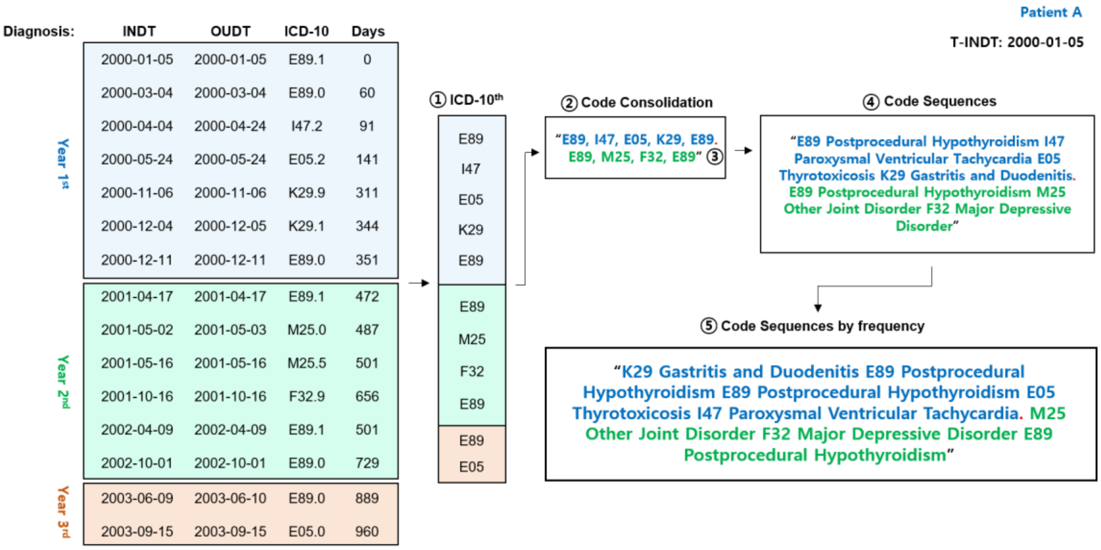
The extracted EMR data from 762,224 individuals comprises various pieces of medical information. Within this dataset, structured information encompasses basic details, physical measurements, visit histories, surgical procedures, medication records, and test results, among others. Unstructured data include diagnostic details and codes (ICD-10th), prescription codes, detailed medication ingredient names, as well as records of Percutaneous Coronary Intervention (PCI) or Coronary Artery Bypass Grafting (CABG) tests.

Data identification was categorized into hospital registration numbers and names. These identifiers were removed to ensure the protection of personal information. This study used ethically pre-approved data and underwent AMC Institutional Review Board (IRB) review (IRB 2021-0303).

Data preprocessing

Structured EMR data, including information such as ages, vital signs, and other relevant details, underwent data cleaning to eliminate outliers and incorrect entries following clinical standards. In the case of unstructured data in coded formats within EMRs, we developed specific preprocessing methods to leverage this data for building a model to predict patient diagnoses and treatments. These preprocessing methods focused on transforming unified code streams encompassing all patient diagnostic codes.

This study utilized ICD-10th codes, a format for recording diagnostic information, to generate 22 highest-level categories. Considering the lowest classification, ICD-10th codes consist of 1,900 [45], and it is common to use only some codes by selecting the most relevant ones. However, we decided to include all ICD-10th codes present in the patient data, including rare codes, to account for all symptom patterns present.



Diagnosis:	INDT	OUDT	ICD-10	Days
Year 1st	2000-01-05	2000-01-05	E89.1	0
	2000-03-04	2000-03-04	E89.0	60
	2000-04-04	2000-04-24	I47.2	91
	2000-05-24	2000-05-24	E05.2	141
	2000-11-06	2000-11-06	K29.9	311
	2000-12-04	2000-12-05	K29.1	344
	2000-12-11	2000-12-11	E89.0	351
Year 2nd	2001-04-17	2001-04-17	E89.1	472
	2001-05-02	2001-05-03	M25.0	487
	2001-05-16	2001-05-16	M25.5	501
	2001-10-16	2001-10-16	F32.9	656
	2002-04-09	2002-04-09	E89.1	501
	2002-10-01	2002-10-01	E89.0	729
Year 3rd	2003-06-09	2003-06-10	E89.0	889
	2003-09-15	2003-09-15	E05.0	960

"E89, I47, E05, K29, E89. E89, M25, F32, E89"

"E89 Postprocedural Hypothyroidism I47 Paroxysmal Ventricular Tachycardia E05 Thyrotoxicosis K29 Gastritis and Duodenitis. E89 Postprocedural Hypothyroidism M25 Other Joint Disorder F32 Major Depressive Disorder"

"K29 Gastritis and Duodenitis E89 Postprocedural Hypothyroidism E89 Postprocedural Hypothyroidism E05 Thyrotoxicosis I47 Paroxysmal Ventricular Tachycardia. M25 Other Joint Disorder F32 Major Depressive Disorder E89 Postprocedural Hypothyroidism"

Figure 6. Overview of preprocessing methods for ICD-10th codes based on patient visits.

Figure 6 illustrates the design of preprocessing methods aimed at creating recipient activities for encoding the entire set of ICD-10th codes, taking into account the visit units in each patients diagnostic code.

First, to leverage the segmental symmetry of ICD-10th codes, we custom-scaled them by rounding the codes to three decimal places (e.g., removing the '1' from diagnosis E89.1).

Second, we consolidated separate codes from each patients diagnostic record to maintain a single code for the classification diagnosed during the year at the visit date (INDT) and discharge date (OUDT). This allowed for the segmentation of short-term and long-term visit data by linking segments within a specific year into a single layer. Cases where patients received fewer than two diagnoses within a year were excluded during this process.

Third, '.' separators were used to preserve the chronological order of various diagnoses and establish relationships in the entire patient visit record. These separators distinguish diagnoses for each patient year.

Fourth, by integrating ICD-10th codes with textual diagnoses, specific diagnostic occurrence histories for each patient were presented in a concise document format, resulting in integrated diagnostic records for each patient within the processed diagnostic records. This specific process generates a code stream updated annually, with a '.' serving as a separator to differentiate each patient's symptom history. In addition, duplicate codes diagnosed within one-year were removed.

Finally, we applied a specific ICD-10th code frequency sorting approach to leverage the context and properties of the BERT Masked Language Model (MLM). This sorting method considered the frequency of ICD-10th code occurrence for all patients who visited AMC Cardiology or Thoracic Surgery. Through this process, one code sequence was generated per patient.

Data Construction

Our methodology also involved fine-tuning a pre-trained BERT model and optimizing disease prediction model to achieve our objectives. To accomplish these objectives, we created two separate cohorts.

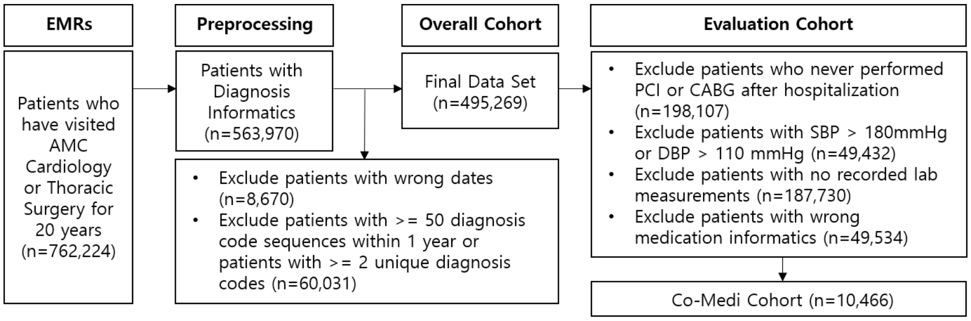

EMRs
Patients who have visited AMC Cardiology or Thoracic Surgery for 20 years (n=762,224)
Preprocessing
Patients with Diagnosis Informatics (n=563,970)
• Exclude patients with wrong dates (n=8,670)
• Exclude patients with >= 50 diagnosis code sequences within 1 year or patients with >= 2 unique diagnosis codes (n=60,031)
Overall Cohort
Final Data Set (n=495,269)
Evaluation Cohort
• Exclude patients who never performed PCI or CABG after hospitalization (n=198,107)
• Exclude patients with SBP > 180mmHg or DBP > 110 mmHg (n=49,432)
• Exclude patients with no recorded lab measurements (n=187,730)
• Exclude patients with wrong medication informatics (n=49,534)
Co-Medi Cohort (n=10,466)

Figure 7. Formation of the patient cohort in the AMC database.

According to Figure 7, we established a cohort comprising a total of 495,269 patients by excluding individuals with no diagnostic records, those with incorrect date formats, or those with an excessive number of diagnostic codes within a year. The dataset consists of individuals and includes code sequences utilizing both diagnostic codes and names. This dataset is employed for the prediction of heart-related diseases, facilitating a direct comparison with the OHE method and the code-embedded model.

In addition, we selectively sampled individuals for this cohort from the overall dataset of 495,269 patients, focusing on those who underwent PCI or CABG surgeries within approximately 7 days of admission, based on the admission date (INDT). We established the 'Co-Medi Cohort,' comprising 10,466 individuals, to assess our model's performance in real-world clinical disease prediction tasks. This cohort incorporates both unstructured EMR data related to diagnoses and prescribed medications, as well as structured data. It also includes essential demographic details such as ages, genders, BMIs, and standardized measurements like vital signs, LDL-C, and Triglycerides. We utilized ICD-10th codes, as well as diagnostic and prescription history in EMRs, to define comorbidities such as Chronic Kidney Disease (CKD), Diabetes Mellitus (DM), Hypertension, and others. The detailed definitions of comorbidities are as follows:

- Chronic Kidney Disease (ICD-10th code, N18) was defined as having an eGFR ≤ 90 before the index date.

- Hypertension (ICD-10th codes, I10-I13, I15) included individuals with a history of prescriptions for beta-blockers, RAAS inhibitors, and patients with a history of using one or more calcium channel blockers.

- Diabetes mellitus (ICD-10th code, N18) was defined as having HbA1c ≥ 6.5%.

The prescription medication-related data is presented in text format, encompassing prescription codes, medication ingredient names, and medication classification codes. Furthermore, we integrated prescription medication-related data to evaluate the embedding method's capability to predict clinical diseases, applying the same preprocessing technique used for diagnostic code sequences.

Models

Data architecture

Using BERT masked language model (MLM)

We selected the BERT MLM model for our study due to the extensive nature of diagnostic code sequences commonly encountered in medical records. Traditional RNN- or CNN-based models struggle with efficient parallel computation in handling these long sequences [46]. BERT, operating on a transformer architecture, features multilayer bidirectional transformers as internal encoders, allowing for the pre-training of advanced linguistic representations [47]. It outperforms GPT, which utilizes a unidirectional architecture, and provides greater potential for training deeper neural networks compared to shallower, bidirectional concatenative structures such as ELMo [48].

In the BERT MLM models, some specific tokens within the input sequences are randomly masked at specified rates, and the models are trained to predict these masked tokens. Through this process, the models utilize two-way parallel computations to learn contextual information and relationships between tokens in the sequences. The models determine the meanings and contextual relevance of each token, understanding the meanings of individual words and their interactions within sentences. During training, the models adjust the weights using gradient descent to predict the token values at each [MASK] location, attempting to minimize the differences between the predicted MASK token values and the actual masked tokens. BERT's bidirectional self-attention mechanisms learn by considering the relationships between all words in sentences, which enables more efficient learning than previous one-way models.

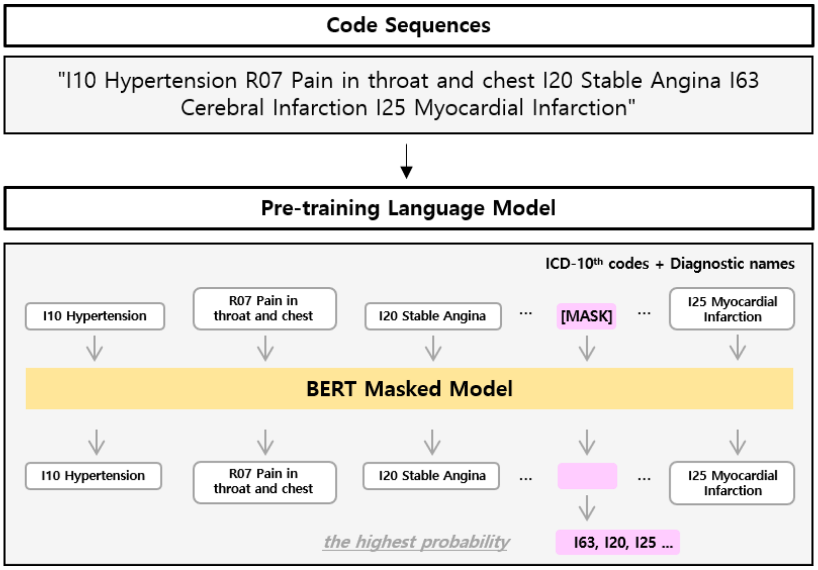

Code Sequences
"I10 Hypertension R07 Pain in throat and chest I20 Stable Angina I63 Cerebral Infarction I25 Myocardial Infarction"
Pre-training Language Model
ICD-10th codes + Diagnostic names
I10 Hypertension
R07 Pain in throat and chest
I20 Stable Angina
…
[MASK]
…
I25 Myocardial Infarction
BERT Masked Model
I10 Hypertension
R07 Pain in throat and chest
I20 Stable Angina
…
…
I25 Myocardial Infarction
the highest probability
I63, I20, I25 …

Figure 8. Visualization of BERT MLM's token masking and predictive learning.

For example, in Figure 8, when the models encounter the masked meaning ‘pain in the throat and chest, hypertension, [MASK] infarction,’ this may be completed with specific heart-related features such as ‘cerebral’ or ‘myocardial.’ This allows the models to determine relationships between various heart-related diagnostic codes and their locations. By employing this approach, the models can help identify detailed characteristics of heart-related conditions and other diagnostic codes associated with them.

These unique aspects of BERT MLM make it well-suited for capturing complex patterns and dependencies in long sequences of diagnostic codes, rendering it a suitable choice for code-embedded models aiming to predict subsequent diseases based on previously recorded patient data.

Code Embedded XGB Model for Heart Disease Prediction

We developed two code-embedded models based on BERT to evaluate the performance of the embedding methods.

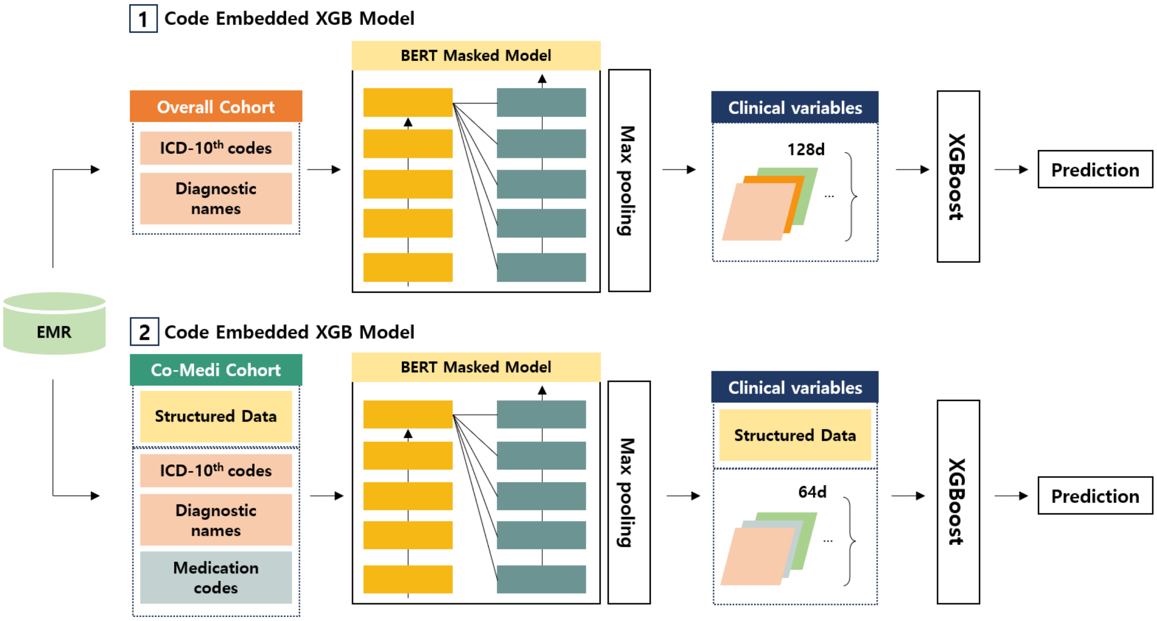


Figure 9. Overview of the code-embedded model architecture.

Figure 9 provides an overview of two models utilizing BERT. The first model, 'Code Embedded XGB Models,' predicts heart-related diseases using unstructured text-format diagnostic datasets from the overall cohort. These models leverage the PLM BERT model, which encodes both ICD-10th codes and diagnostic names, generating dense embedding vectors through max pooling. As a result, the diagnostic data are converted into 128-dimensional vectors and trained using the XGBoost (XGB) [49] model. To optimize the model's performance, we conducted hyperparameter tuning through a random search, utilizing a 30% split of the training dataset for validation. We also applied cross-validation techniques to prevent overfitting and ensure unbiased results.

The 'Code Embedded XGB Model' is specifically designed to predict the occurrence of ten major diagnostic codes associated with heart disease. These codes were selected in consultation with cardiology specialists to ensure the experiment's robustness. The ICD-10th codes related to heart disease used in the heart disease prediction model are as follows:

- C16: Malignant neoplasm of the stomach
- C34: Malignant neoplasm of bronchial tubes and lungs
- C50: Malignant neoplasm of the breast
- G20: Parkinson’s disease
- I21: Acute myocardial infarction
- I48: Atrial fibrillation and flutter
- I50: Heart failure
- I63: Cerebral infarction
- I67: Other cerebrovascular diseases
- J18: Pneumonia

Co-Medi Embedded Model for Major Adverse Cardiovascular Events (MACE)

The second model, the 'Co-Medi Embedded Model,' aims to predict major adverse cardiovascular events (MACE) occurring within one year after percutaneous coronary intervention (PCI) or coronary artery bypass grafting (CABG) procedures [50]. MACE encompasses conditions such as myocardial infarction (MI), unstable angina, all-cause mortality, and ischemic stroke (IS). Here are the specific definitions of MACE:

- Myocardial Infarction (MI) (ICD-10th codes, I21-I23) or Unstable Angina (ICD-10th code, I20) includes patients admitted to hospitals, including those admitted through the emergency room.
- Ischemic Stroke (ICD-10th codes, I63, G45) comprises patients who had undergone at least one brain CT or MR imaging test within 30 days from the date of diagnostic code registration.

Unlike the 'Code Embedded XGB Models,' this model utilizes the 'Co-Medi Cohort' datasets and incorporates not only diagnostic data but also additional information related to baseline clinical variables, comorbidities, and prescribed medications. Further details about these tasks are provided in the 'Methods' section.

Through the BERT MLM model, the combination of the unstructured code sequences of each patients diagnostic and medication code are converted into 64-dimensional continuous numeric vector representations. We use these embedding vectors as input data for the XGBoost (XGB) models to build the ‘Co-Medi XGB Model’. In Addition, we performed hyperparameter tunings to evaluate and compare the MACE prediction performances. We split 30% of the training datasets into validation sets and applied cross-validation techniques to assess the generalizability of the models.

Experimental setup

The code embedded model was developed using Python 3.8 and the Tensorflow DL framework (version 2.6.0), built with a BERT-base architecture. The learning rate for the PLM was fixed at 10^{-3}, and a batch size of 32 was chosen, taking into account the size of the training data and the GPU memory capacity of the Geforce RTX 3090. In addition, the maximum number of tokens was capped at 256, and the vocabulary size was restricted to 7,000, including both the diagnoses and medications. The model fine-tuned with the application of the Adam optimizer, and the optimal epoch was identified based on the corresponding accuracy and loss values.

Model evaluation

To evaluate the performance of the two code-embedding methods, we designed a total of three experiments.

Experiment 1: Comparing the Effects of Code Subsequence Alignment Methods

The first experiment aimed to evaluate the impact of various code subsequence sorting methods on the performance of the BERT MLM. Prior to this experiment, a preliminary investigation was conducted to assess the efficiency of code sequence splitting into different annual increments (e.g., 1, 2, 3, 5, 7 years). These preliminary experiments revealed that the best results for token masking prediction were achieved when the code sequences were split on a yearly basis. Therefore, we standardized the code alignment unit to one year per patient and compared three models based on different diagnostic code sorting techniques.

The first model consisted of a completely random arrangement of code sequences, while the second model was designed to learn diagnostic codes within the sequences in alphabetical order. In contrast, the third model learned diagnostic codes within the sequences based on their frequency of occurrence. We closely monitored loss and accuracy metrics during the training of each model until validation loss no longer improved. This approach facilitated real-time evaluation of the model's learning trajectory and identification of the point of optimal performance.

Experiment 2: Performance Comparison Between OHE and Code-Embedded Methods for Heart-Related Disease Prediction

The second experiment aimed to compare and evaluate the performance difference between OHE and the 'Code Embedded XGB Model' in effectively predicting subsequent diagnoses related to heart diseases. For this purpose, we employed XGBoost to evaluate two models. Specifically, both models utilized the diagnostic dataset from the same overall cohort to construct two separate models: the ‘OHE XGB Model’ and the ‘Code Embedded XGB Model.’ We assessed the prediction performance of code-embedded and OHE using the AUROC (Area Under the Receiver Operating Characteristic curve) metric, which is a commonly used and reliable measure for evaluating model performance [51]. In addition, we used the t-SNE algorithm [52] to visualize and compare each model, measuring the similarity between code sequences in vector space to evaluate the semantic meaning shared among diseases.

Experiment 3: Predicting Major Adverse Cardiovascular Events (MACE) with Co-Medi Embedded Model

The third experiment was conducted to evaluate the performance of the 'Co-Medi Embedded Model' in predicting MACE within one-year for patients who underwent PCI or CABG surgery. This experiment had a similar setup to Experiment 2 but utilized the 'Co-Medi Cohort' dataset. We evaluated the prediction efficiency of code-embedded and OHE using the AUROC metric. This allowed us to compare the performance of each model and confirm their ability to predict actual clinical conditions.

Results

Data characteristics

We have compiled a dataset comprising 495,269 patient records, encompassing diagnostic codes and related diagnostic information, admission and discharge details, as well as baseline patient data. The patient cohort consisted of individuals admitted to the Department of Cardiology or Thoracic Surgery at AMC between January 1, 2000, and December 31, 2019.

The mean ages of patients in the overall cohort were 58.99 years (SD 13.21). The dataset was composed of 46.15% women (228,587 out of 495,269) and 53.84% men (266,680 out of 495,269). The average length of stay (LOS) per visit was 1.94 days (SD 11.26). Primary diagnoses encompassed a range of heart-related diseases, including I10 (hypertension), E11 (diabetes mellitus), E78 (lipid metabolism disorders), R07 (chest pain), I63 (stroke), and others. For a detailed distribution of the diagnostic codes among the target patients.

Table 4. 'Co-Medi Cohort' demographic distribution table for prediction of clinical disease in patients who underwent PCI or CABG.

	Study population (n=10,466) **N(%)**
Mean (SD) Age (years)	**63.30 (10.13)**
Male	**7,840 (74.90)**
Mean (SD) BMI (kg/㎡)	25.01 (2.95)
Mean (SD) SBP (mmHg)	126.62 (18.75)
Mean (SD) DBP (mmHg)	73.44 (11.97)
Mean (SD) HDL-C (mg/dL)	43.99 (11.36)
Mean (SD) Triglycerides (mg/dL)	147.62 (92.79)
Mean (SD) Total cholesterol (mg/dL)	173.04 (43.96)
Mean (SD) LDL-C (mg/dL)	**91.369 (36.66)**
Chronic Kidney Disease	3,111 (29.72)
Diabetes mellitus	3,553 (33.94)
Hypertension	10,247 (97.90)
Metabolic syndrome**	10,138 (96.86)

In patients who underwent PCI or CABG, the mean ages of patients for the prediction of clinical diseases were 63.3 years (SD 10.13), and 74.9% (7,840 out of 10,466) were males. The mean BMI was 25.01 (SD 2.95), and the mean LDL-C level was 91.639 (SD 36.66). Comorbidities included chronic kidney disease (29.72%), diabetes (33.94%), and hypertension (97.9%). The distribution for the evaluation cohort, ‘Co-Medi Cohort,’ are detailed in Table 4.

First result: Performance of BERT MLM according to the partial sequence code sorting schema

We conducted an experiment to evaluate the impact of subcode sequence alignment techniques on the performance of the BERT MLM to find the best-performing model. The experiment involved comparing the loss and accuracy metrics of three distinct models, utilizing BERT MLM throughout the training process.

Table 5. Performance of various BERT models by code sequence alignment method.

Model	Examples of code sequences	Loss	Accuracy
Random model	Z72 Other problems related to lifestyle; I10 Other and unspecified primary hypertension; J98 Other disorders of lung; I20 Angina pectoris unspecified.	0.224	0.964
Alphabetical model	I10 Other and unspecified primary hypertension; I20 Angina pectoris unspecified; J98 Other disorders of lung; Z72 Other problems related to lifestyle.	0.274	0.952
Frequency model	**I10 Other and unspecified primary hypertension; I20 Angina pectoris unspecified; Z72 Other problems related to lifestyle; J98 Other disorders of lung.**	**0.124**	**0.977**

As a result of the experiments, as shown in Table 5, the 'Frequency model,' which sorts based on the frequency of diagnostic codes occurrences, exhibited excellent performance with an accuracy of 0.977. It was followed by the 'Random Model' and the 'Alphabetical Model,' which achieved accuracies of 0.964 and 0.952, respectively.

As a result, we chose the 'Frequency model' as our approach, which achieved accuracies of 0.977.

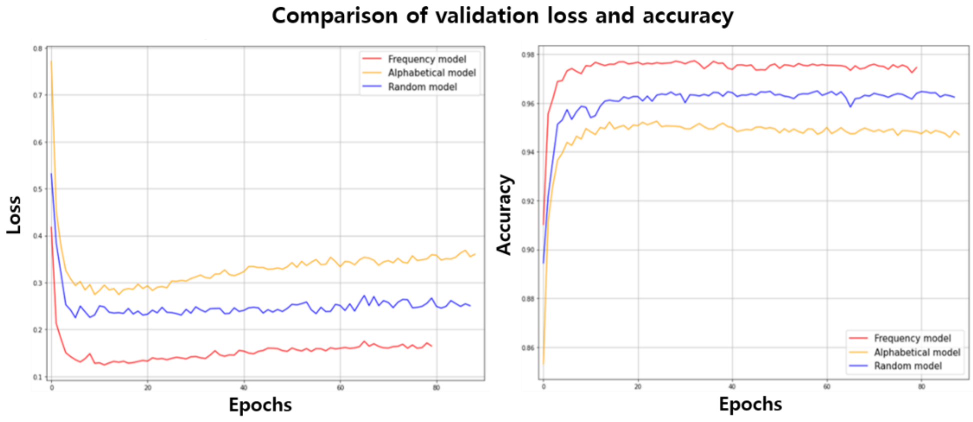


Figure 10. Comparison of loss and accuracy among models trained with different sequence-ordering methods using a dataset of diagnostic code sequences.

In addition, our proposed frequency-based alignment methods demonstrated a loss reduction of approximately 0.1 or more compared to other alignment methods, as illustrated in Figure 10. These experiments confirmed that the way diagnostic codes are sorted significantly impacts the performance of the BERT MLM. These findings suggest that a model prioritizing specific data sorting can effectively minimize loss and enhance accuracy during the training phase.

Performance as a function of model size and embedding pooling method

To evaluate the effectiveness of the selected 'Frequency model,' we conducted model optimization experiments, varying the model's dimensionalities and embedding pooling strategies.

Table 6. Performance of the pooling method in the code-embedded model.

Pooling method	**ICD-10 codes**										**AUC**^{a}
	C16	C34	C50	G20	I21	I48	I50	I63	I67	J18	
Max	0.996	0.996	0.993	0.867	0.986	0.992	0.997	0.981	0.944	0.906	**0.965**
Average	0.982	0.991	0.989	0.855	0.970	0.984	0.984	0.960	0.959	0.917	0.959

[a]AUC: Performance of average AUC values

According to Table 6, the models utilizing max pooling for embedding compression yield a performance metric of 0.965, outperforming the average pooling method with a score of 0.959.

Table 7. Model performance based on differences in parameter sizes and embedding dimensions.

Model	**Size (# Parameters)**	**(# Embedding Dimensions)**	**Accuracy**
Code Embedded model (our methods)	**1,898,590**	**128**	**0.975**
Code Embedded model (normal)	3,921,118	256	0.942
Code Embedded model (medium)	8,359,390	512	0.935
Code Embedded model (large)	13,321,950	768	0.905

Table 7 shows that the code-embedded models achieved the highest accuracy of 0.975 when trained in a 128-dimensional space. This is approximately 0.07 higher than the large models, which had accuracies of 0.905 when trained in a 768-dimensional space, the largest dimension considered.

Therefore, the '128-d Code Embedded Model' were selected as the final comparison models for the final comparison with the OHE-based model.

In conclusion, this suggests that, for certain domain datasets of limited size, increasing model complexity by significantly increasing the number of parameters can negatively impact learning efficiency. The results of our model optimization experiments confirm that model size does not guarantee an absolute improvement in performance.

Second result: Performance of the XGB model in predicting heart-related diseases using OHE and code embedding method

In this section, we evaluated the performances of medical AI models for predicting heart-related diseases using two different approaches, OHE and code-embedded methods. To ensure a balanced comparison, the hyperparameters of the XGB models were kept consistent in both methods.

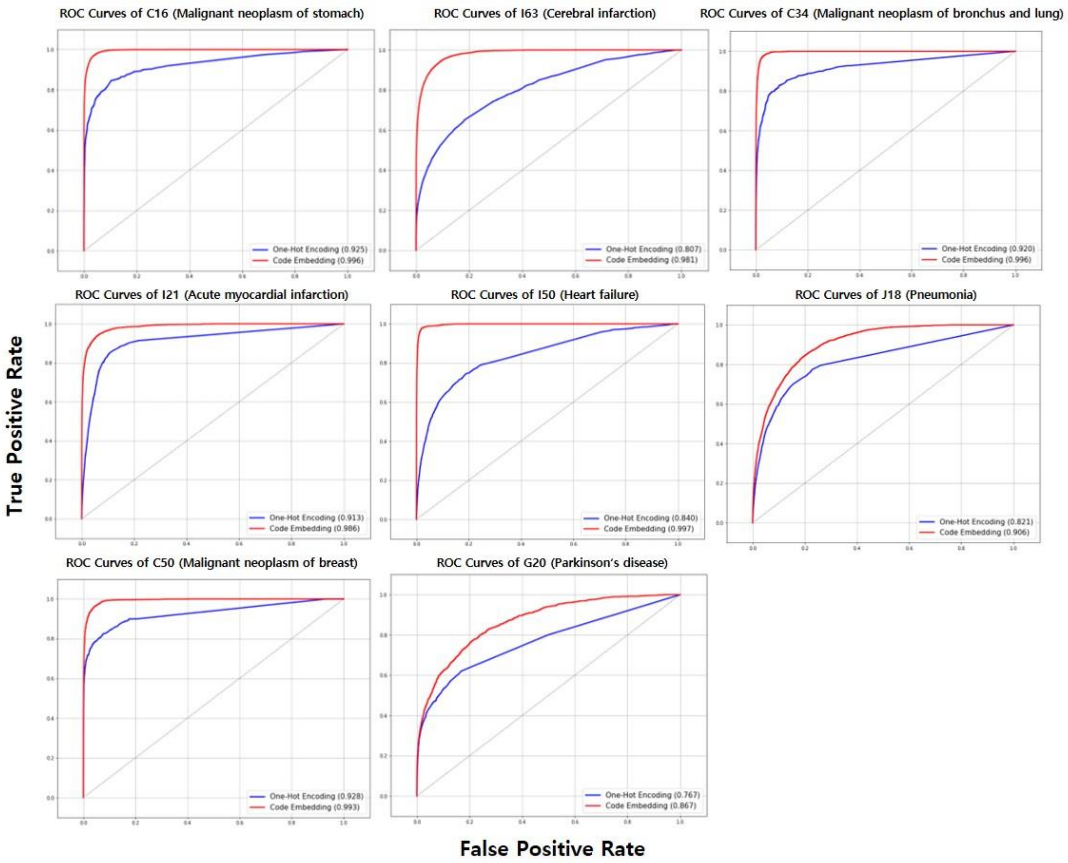


Figure 11. ROC curve comparing the OHE and code-embedded models for predicting heart-related diseases.

According to Figure 11, the XGB models using the code-embedded technique demonstrated excellent performance in predicting I50 (heart failure) with an AUC of 0.997. In contrast, the XGB models using the OHE approach achieved AUC values of 0.840 for the same task, indicating suboptimal performance. Similarly, in the I63 (cerebral infarction) prediction task, the 'Code Embedded XGB Models' achieved high AUC values of 0.981, while the OHE-based models recorded low AUC values of 0.807. Notably, in the G20 (Parkinson's disease) prediction task, both the OHE models and the code-embedded models demonstrated reasonable accuracy. However, the code-embedded models achieved an AUC of 0.867, which was higher than that of the OHE models. Following a comprehensive analysis, the trained code-embedded models, which included diagnostic text, achieved an average AUC performance of approximately 0.96 across 10 heart-related disease prediction tasks. This performance exceeded the average AUC of 0.87 achieved by the OHE models by approximately 0.1.

When combining the results of the two models for heart-related disease predictions, the code-embedded model trained using diagnoses showed superior metrics in both space losses and prediction costs compared to the OHE models. These results also take into account that the OHE models consist of 1,849 dimensions, while our model consists of only 128 dimensions.

Third result: Predicting MACE within one year using medication and diagnostic code embeddings

To evaluate the application of code-embedded methodologies in clinical settings and assess the generalizability of the medical AI models, we conducted experiments predicting the occurrences of real clinical diseases. The models for this evaluation took into account the integration of patient data by incorporating various clinical variables alongside the diagnostic codes. Unlike the 'Code Embedded XGB Model' in Experiment 2, the 'Co-Medi XGB Model' predicts clinical variables by incorporating additional factors such as the patient basic information, prescription codes, medication names, and medication classification codes. Detailed cohort definitions can be found in the 'Methods' section.

Table 8. Comparison of ‘Co-Medi OHE Model’ and ‘Co-Medi Embedded Model’ performance for predicting one-year MACE in patients who underwent PCI or CABG.

Model	#Dimensions	AUC
Basic Clinical Model	**23**	**0.595**
OHE Model	**2,492**	**0.685**
64-d Co-Medi Embedded Model	**87**	**0.749**
128-d Co-Medi Embedded Model	151	0.731
256-d Co-Medi Embedded Model	279	0.732

The performances of the XGB models in predicting MACE within one-year in patients who underwent PCI or CABG are summarized in Table 8. A model that predicted using only basic clinical variables without using both OHE and code-embedded methods recorded an AUC of 0.595. Models using the same parameters as the 'Code Embedded XGB Model' and set to 128 dimensions achieved AUCs of 0.731. We conducted experiments to further adjust the number of dimensions by considering additional clinical variables, and interestingly, the '64-d Co-Medi Embedded Model' demonstrated the highest prediction performances with AUCs of 0.749. In contrast, the ‘OHE Model’ achieved the lowest AUC at 0.685 despite considering 2,492 integrated clinical variables and its high dimensionality.

These experimental results demonstrate that our medical code-embedded algorithm, which integrates prescription medication-related code information alongside diagnostic codes, effectively reduces dimensionality compared to OHE while enhancing prediction performance, even when applied to real-world clinical disease prediction scenarios. In summary, our framework can be applied to real-world medical AI models for clinical disease predictions and offers practical avenues for extending its applicability to various medical prediction models.

Visualization of OHE model and code embedding model

Our goal was to evaluate how effectively each representation technique used in heart-related disease prediction models captured the essential relationships among diseases. To achieve this, we compared the visual representations of diagnostic code relationships between OHE models and code-embedded models using the t-SNE algorithm. These patterns were validated through consultations with clinicians to confirm the appropriateness of the observed relationships among diagnostic codes within disease groups.

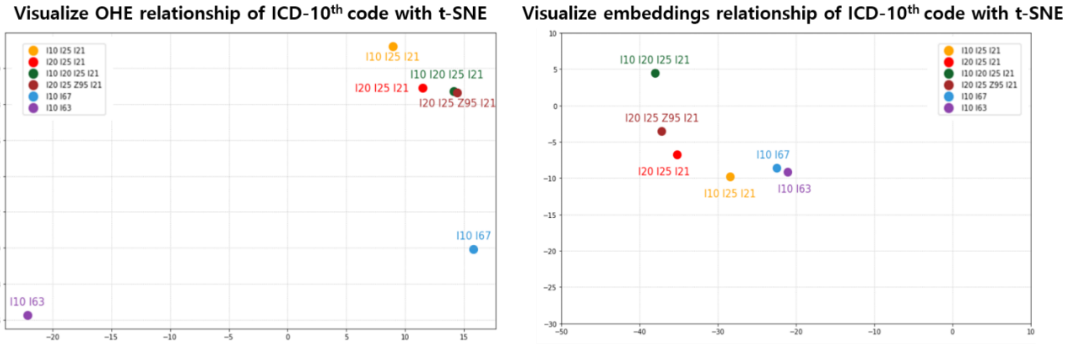


Figure 12. Visualization of ICD-10th code relationships using t-SNE to reduce diagnostic code vectors to two dimensions.

A two-dimensional visualization of diagnostic code vectors can be observed in Figure 12. The visualization results evaluate whether diseases belonging to the relevant clinical groups are clustered based on the classification results. For instance, I10 (hypertension) and I20 (angina) are diagnostic codes commonly associated and contributing to the development of heart-related conditions. In addition, code pairs such as 'I10 I67' (hypertension, other cerebrovascular diseases) and 'I10 I63' (hypertension, cerebral infarction) are generally linked to cerebrovascular diseases through a similar diagnostic sequence. These codes are expected to exhibit strong correlations with concurrently diagnosed codes. Nevertheless, the t-SNE visualization results generated by the OHE models show that these codes are relatively far apart in vector space. In contrast, the code-embedded model more accurately captures significant correlations between patients' diagnostic codes, which appear to have close placements for semantically similar diseases.

Through these two-dimensional vector representations, we confirmed that the code embedded model, utilizing PLM BERT-based contextual embeddings, accurately identifies and represents actual patient diagnostic patterns within diagnostic codes.

Discussion

Findings

The EMRs primarily contain unstructured text data, such as diagnoses, prescriptions, and other medical information, which are often recorded in code format. Currently, these EMR data are mainly integrated into machine learning (ML) models using the OHE method. However, the primary limitation of OHE is its inability to capture the complexity of EMR data, which is predominantly composed of text. Therefore, there is a need for an effective method to extract medical text. In this context, our BERT-based code-embedded model, which applies contextual embedding, has the potential to enhance the clinical prediction performance of medical AI models by understanding the context of EMR text information and effectively integrating medical information in various code formats.

To reflect the diagnostic patterns of actual patients, we collected and processed EMRs from 495,269 patients who visited the Department of Cardiology at AMC between 2000 and 2020. To achieve this, a special preprocessing method was introduced and applied to EMR diagnostic data, considering the frequency of occurrence of diagnostic codes and the yearly diagnostic patterns of each patient.

We have confirmed that models based on the frequency of occurrence of diagnostic codes quickly reduce losses and achieve high accuracy. In addition, we confirmed that the size of the models does not have an absolute effect on performance improvement. We optimized model learning performance by fine-tuning the BERT model and reducing the data dimension.

In an actual heart disease-related follow-up disease prediction experiment, we confirmed that the code-embedded model effectively solves data size and code cardinality problems compared to OHE. We found that despite significantly reducing dimensionality, the ‘Code Embedded XGB Models’ recorded an AUC of 0.96, about 0.1 higher than the ‘OHE XGB Models’. In addition, t-SNE visualizations confirmed that interrelated diagnostic codes were located in similar two-dimensional vector spaces. These findings suggest that our code sequence alignment method better understands important patient information in EMRs through NLP-based context embedding and strengthens models that identify associations with clinical diseases such as diagnostic codes or medications.

To assess the real-world generalization and utility of medical AI models, we achieved the highest performance with an AUC of 0.749 using a joint 'Co-Medi Embedded Model' to predict MACE within one-year for patients who underwent PCI or CABG. This model has the versatility to be applied to various clinical studies by integrating other unstructured EMRs. It can incorporate data such as symptom codes, additional mandatory codes, and medication information to enhance its applicability. Particularly in real-world clinical prediction problems, our embedding method reduces dimensions by about 96.5% compared to OHE and demonstrates an approximately 6% improvement in disease prediction performance. These results emphasize that medical code-embedded management, which integrates multiple data sources, is applicable to various medical prediction models and underscores the potential for risk prediction using realistic real-world datasets.

Limitations and future works

The limitations of this study include the following aspects. First, the models were primarily developed using EMRs of patients visiting the Cardiology Department at AMC, which provide high predictive power for heart-related diseases. Although these approaches enable specialized prediction models for various medical fields and clinical information and form the basis for generating embedding models tailored to specific diseases or conditions, they may somewhat limit the general clinical prediction results of the models. Therefore, future research should focus on developing comprehensive healthcare AI prediction models by leveraging integrated EMRs collected from various healthcare departments and striving for broader applications.

Second, the 'Co-Medi Embedded Models,' designed for predicting MACE within one-year, partially incorporate basic patient information, with a primary focus on diagnostic and medication-related codes. Because the cause of a disease is influenced by many variables, including genetics, lifestyle, environmental factors, underlying conditions, and treatment, our models may have limitations in clinical prediction results. However, an important feature of our models is their scalability, achieved through effective preprocessing and embedding techniques for unstructured EMRs. Therefore, in future research, we aim to develop medical AI models that are more accurate and suitable for realistic medical environments. This entails considering laboratory test results, patients' medical histories, and various other clinical information.

Conclusion

In conclusion, our study proposed a BERT-based code-embedded model developed taking into account the unique challenges presented by the unstructured text format of EMRs. This model improves clinical problem prediction performance by introducing innovative techniques to contextually embed EMR data and reduce code features to lower dimensions. Our model was successful in predicting patients' clinical conditions by encompassing specific codes for diagnoses, medications, and prescriptions. In particular, this model has proven to have superior actual clinical prediction ability compared to models using the existing OHE method, and has been shown to fully understand the complex mechanisms of patient information contained in EMRs.

This approach helps medical professionals make informed decisions, and medical AI prediction models using various EMRs are expected to be of great help in providing important guidance in real-world clinical situations.

Conclusions

This study emphasizes their vital role in adapting to the growing volume of medical data and emerging healthcare technologies, ultimately making a positive impact on disease prediction and treatment for real patients.

Firstly, the study successfully constructed patient cohorts based on Lp(a) levels using EMR data from Korean patients with atherosclerotic cardiovascular disease (ASCVD). It employed various statistical methods to validate the correlation between Lp(a) levels and cardiovascular disease (CVD). This research provides insights into crucial clinical characteristics related to CVD, advancing our understanding of disease prevention and treatment. It underscores the importance of RWE clinical research utilizing EMR data and anticipates that these findings will contribute to the prevention and treatment of CVD.

Secondly, the study proposed a methodology to enhance the performance of medical AI models by extracting EMR data, utilizing actual patient diagnostic history, medication records, and developing a diagnostic code embedding model. It confirmed that a specific sorting order of diagnostic codes can improve the fine-tuning performance of pre-trained language models (PLMs). Moreover, it directly compared code embedding and OHE Encoding approaches to validate the performance of the Code Embedded XGBoost (XGB) model in predicting ten specific diagnostic codes related to heart disease. In conclusion, the code embedding model developed through fine-tuning BERT PLM on frequency-sorted diagnostic codes effectively captured relationships among various diseases and presented nuanced representations of medical information. This research addresses issues related to high-dimensional features in medical datasets and demonstrates the potential to enhance the efficiency of future research based on EMR.

Lastly, this study opens new opportunities for disease prevention and treatment in the medical field and confirms the effective potential of utilizing EMR and medical AI research. Future research endeavors are expected to expand and advance medical AI models based on EMR data, leading to the development of more Pre-trained Language models that utilize extensive EMR text data.

Supplemental Contents

eTable 1. The medication codes for immune checkpoint inhibitors, proton pump inhibitors, and the baseline medication use of study subjects.

Medication 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, captopril, cilazapril, 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, 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
Estrogen/Hormone replacement therapy	chlormadinone acetate, estradiol valerate, hydroxyprogesterone caproate, medroxyprogesterone acetate, megestrol acetate, raloxifene hcl

eTable 2. Diagnosis codes for the assessment of the study subjects' 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) Code for Rare and Intractable Disease Special Calculation Exception: V103 (Human immunodeficiency virus [HIV] disease)

Hypothyroidism	E02 (Subclinical iodine-deficiency hypothyroidism) E03 (Other hypothyroidism) E06.3 (Autoimmune thyroiditis)
Hyperthyroidism	E05 (Thyrotoxicosis [hyperthyroidism])
Liver Disease	B18 (Chronic viral hepatitis) B19 (Unspecified viral hepatitis) K70 ~ K77 (Diseases of liver)

*Rheumatoid arthritis, Psoriasis, HIV were exclusively utilized in the primary analysis results.

eTable 3. The delineation of primary and secondary endpoints pertinent to cardiovascular disease.

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.
Stroke and TIA	*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 ischemic 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 recorded notes for CAG, PCI or CABG procedures. 3. Cardiac enzyme during hospitalization.

eTable 4. Baseline characteristics by Lp(a) quintiles.

N=47818	Quintile 1 (n:9449) [Lp(a) range: 0.3<=lpa<8.3]	Quintile 2 (n:9635) [Lp(a) range: 8.3<=lpa<15]	Quintile 3 (n:9582) [Lp(a) range: 15<=lpa<24.9]	Quintile 4 (n:9583) [Lp(a) range: 24.9<=lpa<43.9]	Quintile 5 (n:9569) [Lp(a) range: 43.9<=lpa<=684]
Mean (SD) Age (years)	61.05 (11.83)	61.54 (11.59)	61.86 (11.51)	61.99 (11.47)	**62.47 (11.24)**
Age ≥65 y (%)	3818	4041	4207	4193	4387
Male	6452	6367	6251	6279	5775
History of ASCVD					
ASCVD subtype					
MI	1620	1885	1864	1789	1728
Angina	2507	2401	2353	2394	2376
Asymptomatic CAD	1079	1143	1103	1176	1280
Ischemic stroke/TIA	551	493	503	529	567

PAD	199	217	208	216	249
Prior PCI/CABG	1874	1827	1787	1681	1624
Smoking status					
Never smoker	4724	4890	4942	4855	5148
Ex-smoker	1860	1833	1650	1731	1588
Current smoker	1711	1598	1602	1579	1413
Mean (SD) BMI (kg/㎡)	24.838 (2.98)	24.711 (2.98)	24.594 (2.96)	24.564 (3.00)	24.376 (3.02)
Baseline Labs					
Mean (SD) SBP (mmHg)	125.3 (19.99)	124.75 (20.16)	125.06 (20.22)	125.36 (21.12)	126.38 (21.69)
Median (Q1-Q3) SBP (mmHg)	124 (111-137)	123 (110-137)	123 (110-137)	123 (110-138)	124 (110-140)
Mean (SD) DBP (mmHg)	73.52 (12.23)	73.3 (12.13)	73.2 (12.35)	73.32 (12.66)	73.42 (12.68)
Median (Q1-Q3) DBP (mmHg)	73 (65-80)	72 (65-80)	72 (65-80)	72 (65-80)	72 (65-80)
eGFR < 60 mL/min/1.73 ㎡	340	237	229	221	273
ACR ≥30mg/g	338	263	302	290	346
Mean (SD) Total cholesterol (mg/dL)	158.66 (42.30)	164 (41.6)	165.54 (41.98)	167.39 (42.4)	172.2 (45.65)
Median (Q1-Q3) Total cholesterol (mg/dL)	156 (128-184)	162 (134-191)	163 (136-191)	165 (138-194)	169 (140-199)
Mean (SD) LDL-C (mg/dL)	87.05 (35.94)	93.77 (35.58)	96.34 (36.21)	99.28 (37.32)	**102.95 (39.16)**
Median (Q1-Q3) LDL-C (mg/dL)	84 (60-109)	91.6 (67.4-116.6)	93.6 (70.2-119)	97 (72-122.2)	99 (74.8-126.8)
Mean (SD) HDL-C (mg/dL)	44.31 (12.98)	43.72 (12.56)	43.68 (12.55))	43.62 (12.4)	44.09 (12.94)
Median (Q1-Q3) HDL-C (mg/dL)	43 (36-51)	42 (35-51)	42 (35-51)	42 (35-51)	42 (35-51)
Mean (SD) Triglycerides (mg/dL)	144.24 (100.47)	137.46 (85.91)	129.55 (73.23)	126.97 (71.84)	127.47 (69.23)
Median (Q1-Q3) Triglycerides (mg/dL)	120 (85-173)	117 (85-165)	112 (82-157)	112 (82-152)	112 (83-153)
Mean (SD) Lp(a) (mg/dL)	5.68 (1.5)	11.40 (1.92)	19.51 (2.83)	32.96 (5.38)	73.99 (30.77)
Median (Q1-Q3) Lp(a) (mg/dL)	5.8 (4.5-6.9)	11.3 (9.7-13.1)	19.4 (17.1-21.9)	35.3 (28.2-37.3)	65.6 (52.7-85.4)

Mean (SD) Level of stenosis					
CAG - Moderate stenosis (50-69%)	14	9	10	7	5
CAG - Severe stenosis (≥70%)	65	64	59	50	69
CCTA – Moderate or Severe	11	*8*	7	7	12
Chronic Kidney Disease	4635	4449	4414	4494	**4884**
Diabetes mellitus	2687	2622	2722	2659	2999
Metabolic syndrome	8081	8148	7988	7939	7910
Hypertension	6932	7139	7181	7197	**7410**
Congestive heart failure	449	446	450	506	**540**
Atrial fibrillation	589	550	537	478	421
Cancer	657	582	650	664	627
Rheumatoid arthritis	21	26	23	22	**30**
Psoriasis	9	6	9	13	8
HIV	2	0	1	0	0
Hypothyroidism	126	136	129	81	115
Hyperthyroidism	59	59	35	30	27
Liver disease	415	301	299	247	208
Lipid lowering treatments before the index date					
Statin					
High intensity	452	413	389	353	470
Moderate intensity	3987	3623	3724	3776	3985
Low intensity	153	157	118	135	184
Ezetimibe	30	17	28	12	28
Other lipid lowering treatments (fibrate, niacin, cholestyramine)	136	110	108	92	79
Aspirin or P2Y12 inhibitor	496	335	292	231	272
Beta-blocker	1843	2148	2299	2424	2467

RAAS inhibitor (ACE inhibitor, ARB, or aldosterone antagonist)	906	986	987	996	1144
Calcium Channel Blocker	467	469	528	565	730
Estrogen/Hormone replacement therapy	106	120	115	119	169
Lipid lowering treatments after the index date					
Statin					
High intensity	720	649	623	573	720
Moderate intensity	6212	6245	6364	6361	6515
Low intensity	215	262	234	262	321
Ezetimibe	92	83	99	89	140
Other lipid lowering treatments (fibrate, niacin, cholestyramine)	325	281	234	247	231
Calendar Year (Index date)					
2001-2005	1751	2521	2607	2860	2991
2006-2010	2180	2966	3082	3157	2874
2011-2015	2222	2107	2025	1893	1742
2016-2020	3296	2041	1868	1673	1962

*Data are number (%) of study population unless stated otherwise

References

1. World Health Organization. (2002). The world health report 2002: reducing risks, promoting healthy life. World Health Organization.
2. Gaidai, O., Cao, Y., & Loginov, S. (2023). Global cardiovascular diseases death rate prediction. Current Problems in Cardiology, 101622.
3. Gaziano, T., Reddy, K. S., Paccaud, F., Horton, S., & Chaturvedi, V. (2006). Cardiovascular disease. Disease Control Priorities in Developing Countries. 2nd edition.
4. 이현종, 배지철, 성기철, 박성근, 전창욱, 류승호, ... & 박정로. (2006). 외견상 건강한 한국인의 혈중 Lipoprotein (a) 분포와 심혈관질환의다른 위험인자와의 연관성. Korean Circulation Journal, 36(2), 150-158.
5. Berg, K. (1963). A new serum type system in man—the Lp system. Acta Pathologica Microbiologica Scandinavica, 59(3), 369-382.
6. Lau, F. D., & Giugliano, R. P. (2022). Lipoprotein (a) and its significance in cardiovascular disease: a review. Jama Cardiology.
7. Kamstrup, P. R., Tybjaerg-Hansen, A., Steffensen, R., & Nordestgaard, B. G. (2009). Genetically elevated lipoprotein (a) and increased risk of myocardial infarction. Jama, 301(22), 2331-2339.
8. Luke, M. M., Kane, J. P., Liu, D. M., Rowland, C. M., Shiffman, D., Cassano, J., ... & Ellis, S. G. (2007). A polymorphism in the protease-like domain of apolipoprotein (a) is associated with severe coronary artery disease. Arteriosclerosis, thrombosis, and vascular biology, 27(9), 2030-2036.
9. Clarke, R., Peden, J. F., Hopewell, J. C., Kyriakou, T., Goel, A., Heath, S. C., ... & Farrall, M. (2009). Genetic variants associated with Lp (a) lipoprotein level and coronary disease. New England Journal of Medicine, 361(26), 2518-2528.
10. Lamon-Fava, S., Marcovina, S. M., Albers, J. J., Kennedy, H., DeLuca, C., White, C. C., ... & Schaefer, E. J. (2011). Lipoprotein (a) levels, apo (a) isoform size, and coronary heart disease risk in the Framingham Offspring Study. Journal of lipid research, 52(6), 1181-1187.
11. O'Donoghue, M. L., Fazio, S., Giugliano, R. P., Stroes, E. S., Kanevsky, E., Gouni-Berthold, I., ... & Sabatine, M. S. (2019). Lipoprotein (a), PCSK9 inhibition, and cardiovascular risk: insights from the FOURIER trial. Circulation, 139(12), 1483-1492.
12. Marcovina, S. M., Albers, J. J., Wijsman, E., Zhang, Z., Chapman, N. H., & Kennedy, H. (1996). Differences in Lp [a] concentrations and apo [a] polymorphs between black and white Americans. Journal of lipid research, 37(12), 2569-2585.
13. Nordestgaard, B. G., Chapman, M. J., Ray, K., Borén, J., Andreotti, F., Watts, G. F., ... & Tybjærg-Hansen, A. (2010). Lipoprotein (a) as a cardiovascular risk factor: current status. European heart journal, 31(23), 2844-2853.
14. Guan, W., Cao, J., Steffen, B. T., Post, W. S., Stein, J. H., Tattersall, M. C., ... & Tsai, M. Y. (2015). Race is a key variable in assigning lipoprotein (a) cutoff values for coronary heart disease risk assessment: the Multi-Ethnic Study of Atherosclerosis. Arteriosclerosis, thrombosis, and vascular biology, 35(4), 996-1001.
15. Yoon, Y. H., Ahn, J. M., Kang, D. Y., Lee, P. H., Kang, S. J., Park, D. W., ... & Park, S. J. (2021). Association of lipoprotein (a) with recurrent ischemic events following percutaneous coronary intervention. Cardiovascular Interventions, 14(18), 2059-2068.
16. Bloomfield, D., Carlson, G. L., Sapre, A., Tribble, D., McKenney, J. M., Littlejohn III, T. W., ... & Pasternak, R. C. (2009). Efficacy and safety of the cholesteryl ester transfer protein inhibitor anacetrapib as monotherapy and coadministered with atorvastatin in dyslipidemic patients. American heart journal, 157(2), 352-360.

17. Gutstein, D. E., Krishna, R., Johns, D., Surks, H. K., Dansky, H. M., Shah, S., ... & Wagner, J. A. (2012). Anacetrapib, a novel CETP inhibitor: pursuing a new approach to cardiovascular risk reduction. Clinical Pharmacology & Therapeutics, 91(1), 109-122.
18. Stein, E. A., Mellis, S., Yancopoulos, G. D., Stahl, N., Logan, D., Smith, W. B., ... & Swergold, G. D. (2012). Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. New England Journal of Medicine, 366(12), 1108-1118.
19. Nordestgaard, B. G., Chapman, M. J., Ray, K., Borén, J., Andreotti, F., Watts, G. F., ... & Tybjærg-Hansen, A. (2010). Lipoprotein (a) as a cardiovascular risk factor: current status. European heart journal, 31(23), 2844-2853.
20. Kim, N. H., & Kim, S. G. (2020). Fibrates revisited: potential role in cardiovascular risk reduction. Diabetes & Metabolism Journal, 44(2), 213-221.
21. Colantonio, L. D., Monda, K. L., Rosenson, R. S., Brown, T. M., Mues, K. E., Howard, G., ... & Muntner, P. (2019). Characteristics and cardiovascular disease event rates among African Americans and whites who meet the Further Cardiovascular Outcomes Research with PCSK9 Inhibition in Subjects with Elevated Risk (FOURIER) trial inclusion criteria. Cardiovascular drugs and therapy, 33, 189-199.
22. Lindh, M., Banefelt, J., Fox, K. M., Hallberg, S., Tai, M. H., Eriksson, M., ... & Qian, Y. (2019). Cardiovascular event rates in a high atherosclerotic cardiovascular disease risk population: estimates from Swedish population-based register data. European Heart Journal-Quality of Care and Clinical Outcomes, 5(3), 225-232.
23. Cegla, J., Neely, R. D. G., France, M., Ferns, G., Byrne, C. D., Halcox, J., ... & Scientific and Research Committee. (2019). HEART UK consensus statement on Lipoprotein (a): A call to action. Atherosclerosis, 291, 62-70.
24. Tsimikas, S., Fazio, S., Ferdinand, K. C., Ginsberg, H. N., Koschinsky, M. L., Marcovina, S. M., ... & Liu, L. (2018). NHLBI working group recommendations to reduce lipoprotein (a)-mediated risk of cardiovascular disease and aortic stenosis. Journal of the American College of Cardiology, 71(2), 177-192.
25. Terpening, C. M. (2019). A call for more complete reporting of cardiovascular death. Circulation, 140(11), 887-888.
26. Johannesen, C. D. L., Langsted, A., Mortensen, M. B., & Nordestgaard, B. G. (2020). Association between low density lipoprotein and all cause and cause specific mortality in Denmark: prospective cohort study. Bmj, 371.
27. Kim, H. S., Lee, S., & Kim, J. H. (2018). Real-world evidence versus randomized controlled trial: clinical research based on electronic medical records. *Journal of Korean medical science*, *33*(34).

28. Hannan, T. J. (1996). Electronic medical records. *Health informatics: An overview*, *133*.
29. Zhao, C., Jiang, J., Xu, Z., & Guan, Y. (2017). A study of EMR-based medical knowledge network and its applications. *Computer methods and programs in biomedicine*, *143*, 13-23.
30. [Deep text summarization] Moradi, M., Dorffner, G., & Samwald, M. (2020). Deep contextualized embeddings for quantifying the informative content in biomedical text summarization. *Computer methods and programs in biomedicine*, *184*, 105117.
31. 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*.
32. Si, Y., Du, J., Li, Z., Jiang, X., Miller, T., Wang, F., ... & Roberts, K. (2021). Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. *Journal of biomedical informatics*, *115*, 103671.
33. Xiang, X., Duan, S., Pan, H., Han, P., Cao, J., & Liu, C. (2020, December). From One-Hot Encoding to Privacy-Preserving Synthetic Electronic Health Records Embedding.

In *Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies* (pp. 407-413).

34. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
35. Nguyen, D., Luo, W., Venkatesh, S., & Phung, D. (2018). Effective identification of similar patients through sequential matching over ICD code embedding. *Journal of medical systems*, *42*, 1-13.
36. Choi, Y., Chiu, C. Y. I., & Sontag, D. (2016). Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, *2016*, 41.
37. 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.
38. Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, *304*, 114135.
39. Peng, Y., Yan, S., & Lu, Z. (2019). Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.
40. Ji, S., Hölttä, M., & Marttinen, P. (2021). Does the magic of BERT apply to medical code assignment? A quantitative study. *Computers in Biology and Medicine*, *139*, 104998.
41. Gao, S., Alawad, M., Young, M. T., Gounley, J., Schaefferkoetter, N., Yoon, H. J., ... & Tourassi, G. (2021). Limitations of transformers on clinical text classification. *IEEE journal of biomedical and health informatics*, *25*(9), 3596-3607.
42. 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), 1-13.
43. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
44. Ahn, I., Na, W., Kwon, O., Yang, D. H., Park, G. M., Gwon, H., ... & Kim, Y. H. (2021). CardioNet: a manually curated database for artificial intelligence-based research on cardiovascular diseases. *BMC Medical Informatics and Decision Making*, *21*, 1-15.
45. O'malley, K. J., Cook, K. F., Price, M. D., Wildes, K. R., Hurdle, J. F., & Ashton, C. M. (2005). Measuring diagnoses: ICD code accuracy. *Health services research*, *40*(5p2), 1620-1639.
46. Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.
47. 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*.
48. Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512*.
49. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1-4.
50. Kwon, O., Na, W., Kang, H., Jun, T. J., Kweon, J., Park, G. M., ... & Kim, Y. H. (2022). Electronic Medical Record–Based Machine Learning Approach to Predict the Risk of 30-Day Adverse Cardiac Events After Invasive Coronary Treatment: Machine Learning Model Development and Validation. *JMIR Medical Informatics*, *10*(5), e26801.

51. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, *1*(1), 18.

52. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

국문 요약

최근 헬스케어 관련 기술 발전으로 인하여 의료기관에서 수집되는 임상 데이터의 양이 급격히 증가하고 있다. 전자의무기록(EMR)은 임상 데이터 중 하나로, 환자의 다양한 진료 기록을 포함하고 있다. 이러한 데이터는 개인정보 보호를 위해 가명화 또는 익명화 과정을 거쳐 비식별화되어, 의료 분야의 인공지능 모델 개발에 활용되고 있다.

또한, 질병의 복잡한 메커니즘을 이해하고 기존의 방식으로는 치료하기 어려웠던 질병에 대한 치료법을 개발하는 RWE(Real-World Evidence) 연구에 중요한 역할을 하고 있다. RWE 연구는 실제 환자 데이터를 기반으로 질병과 약물, 부작용 등의 상태와 임상적 요소 간의 관계를 명확히 파악하고 환자의 위험을 예측하는 데 기여한다. 따라서, 대규모 EMR 데이터를 활용한 후향적 연구는 다양한 위험 예측 연구를 수행하고 의료 분야에서 RWE 로 활용될 수 있다.

특히 심혈관 질환(CVD)은 전 세계적으로 주요 사망 원인 중 하나로 인식되며, 높은 Lipoprotein(a) 수치는 CVD 관련 사건의 위험을 증가시키는 주요 요인 중 하나로 알려져 있다. 심혈관질환은 여러 동반질환을 동반하는 급성 및 만성 질환 중 하나로, 지속적이고 적극적인 관리가 필요하다. 따라서 한국에서 고위험 ASCVD 과거력을 가진 환자를 대상으로 EMR 을 활용하여 Lp(a)의 수치에 따른 임상적 특성과 심혈관 결과를 추정하는 연구를 수행하고자 한다.

또한, 이러한 임상 연구를 위해 EMR 내 수치 데이터를 보편적으로 사용되는 One-Hot Encoding(OHE) 방식으로 처리하는 것이 일반적이다. 그러나, EMR 데이터는 주로 비정형 텍스트 데이터로 기록되어 있어 의학 텍스트에서 유용한 정보를 추출하는 작업에 한계가 있다. 최근에는 NLP(Natural Language Processing) 기술을 활용하여 텍스트 데이터를 word embedding 방식으로 처리하고 이를 인공지능 모델에 학습시키는 방법이 주목받고 있다. 그럼에도 불구하고, 기존의 연구에서는 환자의 진단 패턴을 고려하지 못하는 한계가 있다. 따라서 우리는 진단 코드 embedding 방법을 사용한 모델을 개발하여 환자의 진단 패턴을 효과적으로 반영하고 이를 통해 의료 인공지능 모델의 성능을 향상시키는 방법을 제안한다. 이는 의료 연구 및 환자 치료에 새로운 가능성을 제공할 것으로 기대한다.

위와 같은 EMR 을 활용한 RWE 임상 연구와 의료 인공지능 모델의 개발 및 활용을 위해 다음 두 가지 연구를 계획하였다.

첫 번째 챕터에서는, 실제 병원 EMR 데이터를 활용하여 실제 환자 데이터를 기반으로 질병 메커니즘을 파악하고 환자의 위험을 예측하는 데 도움을 주는 RWE 연구를 수행하는 것을 목표로 한다. 한국의 ASCVD 환자를 대상으로 EMR 데이터를 추출한다. 심혈관질환과 관련된 임상 데이터는 기본적으로 외래 및 입원 데이터뿐만 아니라 심장초음파 및 뇌졸중 서식지와 같은 다양한 특수 검사 기록을 포함한다. 이러한 다양한 정형 데이터와 비정형 데이터를 통합하여 Lp(a) 수치별로 그룹화한 대상의 코호트를 구축한다. 익명화 된 데이터를 추출하고 임상적으로 수용가능한 기준에 따라 이상치 및 오류 데이터를 제거하였으며, 자연어처리 기법을 활용하여 문장 형태의 검사 결과 기록지 등의 비정형 데이터를 구조화하였다. 임상적 논의를 거쳐 통계적으로 유의미한 임상 변수를 탐색하였으며, 대상 환자군의 특성과 위험 요인 및 MACE

event 발생률 간의 상관 관계를 생존 분석과 RCS curve 등의 다양한 통계적 기법을 활용하여 검증하였다. 심혈관질환으로 입원한 환자들의 Lp(a) 수치와 MACE 재발생 가능성을 예측하였고 이를 통해 질병의 메커니즘을 파악하고 실제 환자의 위험 요인을 예측하는 RWE 임상 연구를 수행하였다.

두 번째 챕터에서는, 환자의 진단 패턴을 효과적으로 반영하는 진단 코드 임베딩 방법을 개발하여 의료 인공지능 모델의 성능 향상을 목표로 한다. EMR 내 환자의 내원 및 퇴원 기록과 진단 코드, 진단명, 진단 일자 등의 데이터를 추출하여 특성 순서를 반영한 코드 시퀀스를 생성하였다. 생성한 코드 시퀀스 데이터셋을 BERT 기반의 PLM model 을 활용하여 임베딩 차원을 최적화한 모델을 개발하였다. 이후 심장 질환과 관련된 10 가지 특정 진단 코드를 예측하는 문제를 설계하여 OHE xgb 모델과 비교하여 성능이 향상되는 것을 검증하였다. 또한 차원 축소 시각화 기법을 사용하여 진단 코드 임베딩 방법을 사용한 우리의 접근 방식이 의료 데이터 내에서 중요한 진단 코드의 복잡한 관계를 효과적으로 포착하는 것을 확인하였다. 실제 환자의 처방 및 진단 정보와 ICD-10 코드로 표준화된 질병 진단 데이터를 사용하여 데이터를 사용한 말뭉치를 구성하고 이를 PLM(Pre-trained Language model) BERT 에 학습시켜 패턴을 고려한 의료 인공지능 모델 개발 연구를 수행하였다.

결과적으로, 우리는 EMR 데이터를 활용하여 심혈관 질환 과거력을 갖는 환자들을 대상으로 Lp(a) 수치와 관련된 임상 코호트를 구축하고, 임상 변수와 재발성 MACE 발생 간의 관계를 다양한 통계적 기법을 사용하여 검증하였다. 또한, EMR 데이터를 활용하여 BERT 기반의 저차원 embedding 모델을 개발하고, OHE 모델과 비교하여 심장 질환 관련 질병 예측 성능을 향상시켰다.

본 연구는 EMR 을 활용한 RWE 임상 연구 및 의료 인공지능 모델 개발의 적용은 질병 원인과 메커니즘을 이해하여 실제 환자의 위험 요소 예측을 위한 의료진의 진단 및 치료 결정 과정에 도움을 줄 수 있다. 또한 이러한 기술적 접근을 통해 EMR 을 활용한 응용 프로그램에서 미래 임상 연구의 효율성 향상에 긍정적인 영향을 줄 것으로 예상한다. 마지막으로, 우리는 EMR text 데이터를 기반으로 의료 분야에 특화된 Pre-trained Language Model (PLM) 개발을 향후 목표로 설정하고 있다. 결론적으로, 본 연구의 확장은 의료 임상 연구와 환자 진단 및 치료 분야에서 새로운 기회가 열릴 것으로 기대한다.