creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

연세대학교
YONSEI UNIVERSITY

Development of machine learning-based model to predict cardiovascular disease in patients at risk using healthcare big data

Shinjeong Song

Department of Medicine

The Graduate School, Yonsei University

연세대학교
YONSEI UNIVERSITY

연세대학교
YONSEI UNIVERSITY

Development of machine learning-based model to predict cardiovascular disease in patients at risk using healthcare big data

Directed by Professor Hyuk-Jae Chang

The Doctoral Dissertation
submitted to the Department of Medicine,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Medical Science

Shinjeong Song

December 2023

연세대학교
YONSEI UNIVERSITY

This certifies that the Doctoral Dissertation of Shinjeong Song is approved.

Thesis Supervisor : Hyuk-Jae Chang

Thesis Committee Member#1 : Tae Hyun Kim

Thesis Committee Member#2 : In Hyun Jung

Thesis Committee Member#3: Yeonyee E Yoon

Thesis Committee Member#4: Jimin Sung

The Graduate School
Yonsei University

December 2023

연세대학교
YONSEI UNIVERSITY

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my gratitude to the many people who helped me to complete my thesis.

I would like to express my sincere gratitude to my advisor, Prof. Hyuk-Jae Chang, who guided my research and provided me with careful guidance until the completion of my thesis. I would also like to express my sincere gratitude to thesis committee members, Professors Tae hyun Kim, In Hyun Jung, Yeonyee E Yoon, and Jimin Sung for their generous guidance during the thesis review process. I am indebted to Gaeun Kim, Taeyoung Kim, Dongyup Shin, Sunhee Kim, Jeongho Park and Jaehyung Bae for their great help during the modeling process. My family deserves endless gratitude: my children Jiwoo Jeong and Yeonwoo Jeong for loving mommy without any conditions, and my parents for their love and wisdom that guided me. I would also like to thank my grandfather, a pastor whom I will always admire and miss, for encouraging me to pursue the path of an academic. I dedicate this doctoral degree to my husband, who has consistently stood by my side throughout our shared journey as physicians, expressing my profound gratitude, appreciation, and love for his unwavering encouragement and support during my career, graduate school, and thesis writing. Last but not least, I thank God for his abundant, abounding grace over my life.

연세대학교
YONSEI UNIVERSITY

<TABLE OF CONTENTS>

연세대학교
YONSEI UNIVERSITY

LIST OF FIGURES

연세대학교
YONSEI UNIVERSITY

LIST OF TABLES

연세대학교
YONSEI UNIVERSITY

연세대학교
YONSEI UNIVERSITY

ABSTRACT

Development of machine learning-based model to predict cardiovascular disease in patients at risk using healthcare big data

Shinjeong Song

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Hyuk-Jae Chang)

The rise in cardiovascular disease worldwide is causing enormous social and economic costs. Accordingly, the field of precision medicine aims to improve care through personalized prediction and prevention. In South Korea, we have health insurance claims data covering almost every citizen, which provides all the information about healthcare utilization behavior. Health insurance users can access their data through a simple authentication process. This data can be used to predict their personalized risk factors. Recently, bidirectional encoder representations from transformers (BERT) and related models have achieved tremendous success in the natural language processing domain. We adapt the BERT framework originally developed for the text domain to the structured HIRA data. The study aimed to predict cardiovascular diseases in subjects at risk (newly diagnosed metabolic diseases; hypertension, diabetes, hyperlipidemia) using health insurance claims data and BERT. Each disease was assigned to the training, validation, and test sets in the ratio of 7:2:1 through data augmentation. Patients' diagnoses and prescribed medications were embedded as input sequences, and age was used for positional encoding to distinguish visits. The model's predictive ability was evaluated by measuring the area under curve (AUC).

In each group of patients diagnosed with hypertension, diabetes, and dyslipidemia, BERT achieved mean AUC areas of 97.9%, 97.8%, and 97.8%, respectively. We found that the

연세대학교
YONSEI UNIVERSITY

top-ranked conditions for self-attendance were hypertension, diabetes, dyslipidemia, and diagnoses and medications that are more common in older adults.

BERT performs good cardiovascular diseases prediction using only diagnosis names and medication prescriptions on a relatively small training dataset. This study suggests that BERT can be used to advance personalized predictive healthcare models and patient care.

Key words : bert, machine learning, metabolic disease, cardiovascular diseases, hypertension, diabetes, dyslipidemia

연세대학교
YONSEI UNIVERSITY

Development of machine learning-based model to predict cardiovascular disease in patients at risk using healthcare big data

Shinjeong Song

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Hyuk-Jae Chang)

I. INTRODUCTION

Cardiovascular disease (CVD), which is the leading cause of death globally, has rapidly increased in public health all over the world.[1] This has resulted in significant social and economic costs. Patients at high risk for CVD can be identified by prediction models that use risk stratification. The field of precision healthcare aims to improve the provision of care through precise and personalized prediction, prevention, and intervention.

Traditionally, risk factors used to predict cardiovascular events include systolic blood pressure, diastolic blood pressure, glycated hemoglobin, cholesterol levels, family history, smoking history, etc. These risk factors for cardiovascular events have not been studied in a national cohort or shared data in Korea. In addition, tools for patients to assess or identify their risk of future cardiovascular events are difficult to access and the interpretation of the results by the general public is very limited. Therefore, it would be meaningful to show disease prediction results using only information on healthcare utilization behaviors that is accessible and shared by everyone.

In Korea, there are Health Insurance Review and Assessment (HIRA) Service, which

연세대학교
YONSEI UNIVERSITY

reviews the claims, assesses the quality of care provided, and evaluates adequacy for healthcare services. HIRA database includes information about healthcare utilization behavior on diagnoses, procedures (examinations), prescription records, visit dates, and demographic characteristics almost all citizens. Health insurance users can access their data through a simple authentication process, and researchers can obtain complete data for research purposes, which is very useful for research purposes or to use as a predictor of individual risk factors for users.

However, this HIRA database is claims data, like statements. Unlike traditional risk factors, it is not numeric but characterized by ICD codes and medication codes, and there is a lot of information in the time series, which limits the development of models using traditional risk factor prediction statistical methods. It is well known that in recent years, advances in deep learning (DL), a subfield of machine learning (ML), have led to significant progress toward personalized predictions in cardiovascular medicine, radiology, neurology, dermatology, ophthalmology, and pathology.[2-5]

The remarkable success of DL in these applications can be attributed not only to advancements in DL algorithms but also to the substantial influx of extensive multimodal biomedical data. These datasets include, among others, electronic health records (EHR)[6], which have played a pivotal role in supporting the development and effectiveness of DL models in the medical domain. With the increasing adoption of electronic health records (EHR) systems in many countries, linking data from tens of thousands of patients over the years, there has been a lot of development on how to use this textual medical information to make predictions using machine learning.

Information about a patient's healthcare utilization behavior, such as multiple outpatient

연세대학교
YONSEI UNIVERSITY

visits or hospitalizations and the medical procedures and medication types associated with them, can generate thousands of data points, while a diagnosis can be a single disease code, making the volume of data suitable for applying ML models and vice versa. Large-scale EHRs therefore provide an unparalleled source of insights and a unique data source for training ML models that require large amounts of data. DL models are gaining popularity in EHR research due to their success in a variety of applications. Various DL approaches [8-10] have been shown to provide good results compared to widely used feature extraction and transformation methods for predicting various diseases from EHR data. In addition, CNN, RNN, and LSTM models have been proposed to account for the complexity of EHRs, such as irregular visit intervals and event sequences. [11-14] Transfer learning was developed to address pre-training some representation on a large unannotated dataset, then training it on a large dataset, and then further tuning it to guide other tasks. [15] A recent trend in transfer learning is the use of self-supervised learning on large general datasets: learning is used to derive a general-purpose pre-trained model that captures the intrinsic structure of the data, which can then be applied to specific tasks on specific datasets through fine-tuning. This pre-training fine-tuning paradigm has proven to be very effective in natural language processing (NLP)[16-20] and more recently in computer vision. [21,22] The bidirectional encoder transducer representation (BERT) is one of the most widely used models for processing sequential inputs such as text and has many variants. [20,23-29] BERTs have also been adopted in the clinical domain [23,24,30] and have been trained on clinical NLP tasks and clinical texts only. Through fine-tuning, they can be used for specific purposes on specific datasets.

연세대학교
YONSEI UNIVERSITY

Therefore, we aimed to predict cardiovascular event in patients at risk; with new-onset metabolic diseases such as hypertension, diabetes, and dyslipidemia using national health insurance claims data represented by ICD codes and drug codes via a BERT model.

연세대학교
YONSEI UNIVERSITY

II. Materials and Methods

1. Study design

The period from 2007 to 2010 was used as a wash out period, and people who were newly diagnosed with hypertension, diabetes, and dyslipidemia between 2011 and 2020 were defined as patients at risk. Among the patients at risk, we divided them into those diagnosed with cardiovascular diseases according to the operational definition (positive) and those without (negative) and trained them with a machine learning model (Bidirectional Encoder Representations from Transformers model) to develop a prediction model. This study was approved by the Institutional Review Board of Yonsei University Severance Hospital.

2. Data.

South Korea has a universal healthcare coverage system, with the National Health Insurance covering approximately 98% of the total South Korean population. The Health Insurance Review and Assessment Agency's claims data covers 46 million patients per year, or 90% of South Korea's population as of 2011, and includes claims from approximately 80,000 healthcare providers across the country. HIRA's claims data includes patients' diagnoses, treatments, procedures, surgical histories, and prescription drugs, making it a valuable resource for healthcare research.

Due to the nature of these HIRA data, understanding the complex structure and large volume of claims data requires significant effort from researchers, so HIRA has developed validated patient sample data from five organizations.

The patient sample is a stratified random sample drawn from HIRA's claims data. The sample size was carefully calculated and drawn on a yearly basis to be representative of Korean patients' sociodemographic characteristics, diagnoses, and prescribed medications.

연세대학교
YONSEI UNIVERSITY

[31] For this study, we were provided with a dataset of 200,000 patients each with hypertension, dyslipidemia, and diabetes directly from the HIRA. Since the amount of data for machine learning was relatively small, we performed augmentation, and after augmentation, we divided the training, validation, and test sets in a 7:1:2 ratio to train the model.

The study sought to predict the development of cardiovascular events in a subject at risk, so the definition of 'subject at risk' was those with newly diagnosed hypertension, diabetes, or dyslipidemia. The operational definitions of hypertension, diabetes, and dyslipidemia are as follows, followed by a list of cardiovascular diseases that are considered complications of each condition. If there were multiple cardiovascular events during the follow-up period, the time point was defined based on the first cardiovascular event.

연세대학교
YONSEI UNIVERSITY

Table 1. Operational definition of the comorbidities

ICD 10 code (1)	Medication (2)	Number of diagnoses (3)	Diagnostic test or treatment (4)	Combination
Subjects at risk				
Hypertension				
I10.x–I13.x, I15.x	Anti-hypertensive drugs	Admission ≥1 or outpatient department ≥1		1+2+3
Diabetes				
E11.x – E14.x		outpatient department ≥2		1+2
E11.x – E14.x	Antidiabetic agent	Admission ≥1 or outpatient department ≥1		1+2+3
Dyslipidemia				
E78.x	Lipid-lowering agent	Admission ≥1 or outpatient department ≥1		1+2+3
Cardiovascular diseases				
Coronary artery disease				
I21-I22		Admission or outpatient department	CAG or CAG with PTCA or Coronary	1+3 or 4

연세대학교
YONSEI UNIVERSITY

	≥1	artery bypass surgery	
Ischemic cerebrovascular disease			
I63-I64	Admission or outpatient department ≥1		1+3
Hemorrhagic cerebrovascular disease			
I60-I62	Admission or outpatient department ≥1	Transfusion	1+3+4
TIA			
G45	Admission or outpatient department ≥1		1+3

3. Pre-processing of data for machine learning model

The data preprocessing process consists of 1) data cleansing, 2) label definition, 3) data augmentation, 4) vocabulary construction, and 5) input data preprocessing.

1) Data cleaning is to merge five tables of billing data and reorganize them into statement units, each of which is as follows. T200: Statement general details, T300: Treatment details, T400: Prescribed diseases, T530: Prescription details, T310: Death information. Drug codes occur multiple times per statement in the T530 table depending on the number of drugs prescribed, so they were reconstructed by grouping them by statement using the separator '|'. Variable names, meanings, and examples are shown in the supplementary table.

연세대학교
YONSEI UNIVERSITY

Table 2. Data cleansing examples

Variables	Definition/Meaning	Example
JID	Provider Number (De-identified)	11111111
MID	Statement number	100001
SEX_TP_CD	Gender (1: Male, 2: Female)	1
PAT_BTH	Patient's date of birth	19660518
RECU_FR_DD	Treatment Start Date	20120705
FOM_TP_CD	Type code (031:Outpatient, 021:Hospitalization, etc.)	031
MAIN_SICK	Main diagnosis code (*KCD code)	I109
SUB_SICK	Sub diagnosis code (KCD code)	E785
GNL_CD	Drug generic name code (main ingredient code) Separator: '|'	123908ACS|262500ATB|427800ACH
VST_DDCNT	Number of inpatient days	0

연세대학교
YONSEI UNIVERSITY

2) Label definition - The subjects of the billing data used to train the label definition model are divided into positive data that develop the disease during the entire data period and negative data that do not develop the disease. Based on the date of the subject's latest statement, we organize the labels in the form of classification by indicating the year in which the specific disease to be predicted develops and labeling it as negative (not developed within the data period) or a specific year of development (positive). Since the data was obtained from 2011 to 2020, 11 classes are generated, where 1 means no disease occurred within 10 years, 2 means disease occurred in year 1, and 11 means disease occurred in year 10.

No.	Negative	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
1	1	0	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0
6	0	0	0	0	0	1	0	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0	0
8	0	0	0	0	0	0	0	1	0	0	0
9	0	0	0	0	0	0	0	0	1	0	0
10	0	0	0	0	0	0	0	0	0	1	0
11	0	0	0	0	0	0	0	0	0	0	1

Figure 1. Label definition. No. 1 means 'no disease within 10 years' and is classified as negative data. No. 2 means "disease occurred in year 1" and is positive data, and in the same order, No. 11 means "disease occurred in year 10" and is classified as positive data.

연세대학교
YONSEI UNIVERSITY

3) Data augmentation - In general, it is known that deep learning models are best suited for large-scale data, and the more data you have, the more potential you have to improve performance. Since the number of dataset is relatively small for training a deep learning model, we tried to compensate for it through data augmentation. In addition, we designed the model in the form of a classification to predict the year of disease onset and one of the important factors in a classification model is the ratio between classes. It is best if the model learns each class evenly, but there may be a class imbalance due to a small number of disease incidence data, or the disease incidence data being divided by year. To compensate for this, we augmented the positive data. Diagnosed subjects (positive) were augmented to generate annualized diagnosis incidence data by

truncating the most recent data by one year based on the first diagnosis treatment start date. Subjects who were not diagnosed (negative) were labeled the same because they were negative, but the data were augmented to create multiple data from a single person by truncating the data by one year based on the last date of care.

The illustration (Figure 2) of the positive data augmentation process is an example of diabetes, where the light blue area on the left is the period used as input data for the model, and [-ny] means the data (statements) corresponding to year n as of the first diagnosis date. If all the data before the diagnosis is used, the label will be the first-year occurrence because the diagnosis is made within one year from the last data, but if the data in the -1y period is not used, the first diagnosis is made in the second year from the last data, so the label is applied as the second year occurrence, and in this way, the positive data is increased by excluding the intermediate data by one year. While positive data has a reference date of the first diagnosis, negative data, i.e., data of subjects who are not diagnosed with the target disease within the time period, does not have a reference date, so we set the last date of the

연세대학교
YONSEI UNIVERSITY

data as the reference date. Since negative data are not diagnosed during the entire time period, the label is always negative regardless of how the data is truncated. We used the historical data exclusion method (Figure 3) and to control the number of voice data, we did not use all augmented data but randomly sampled some of them.

Figure 2. The process of augmenting positive data. In the process of augmenting positive data, the light blue part on the left is the period used as input data for the model, and [-ny] means the data (statements) corresponding to year n as of the first diagnosis date. If all of the data prior to diagnosis is used, the label will be a year 1 occurrence because the diagnosis is made within one year of the last data. If we disable the data in the -1y interval, the first diagnosis will occur in the second year after the last data, so we label it as a second-year occurrence. In this way, we set the label by increasing the positive data by excluding the intermediate data by one year.

연세대학교
YONSEI UNIVERSITY

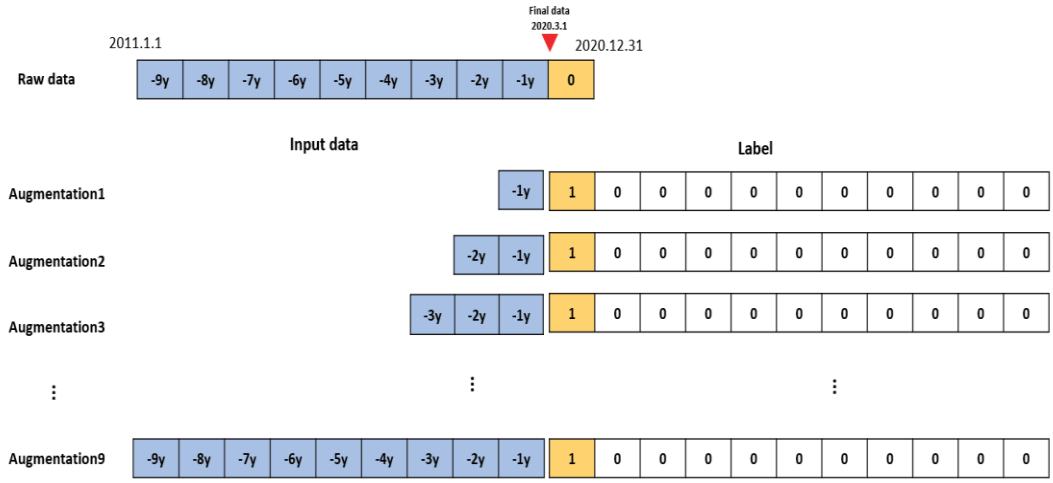

Figure 3. The process of augmenting negative data. Negative data is not diagnosed within the entire time period, so no matter where you cut the data, the label is always negative. We created augmented data by excluding historical data. To control the number of negative data, we did not use all of the augmented data but randomly sampled some of it.

4) Dictionary - Configure a dictionary for mapping disease codes and drug codes in character (String) type to numbers (vector). Main or sub-diagnosis codes consist of 5 or more digits, and to reduce the size of the dictionary, a 3-unit classification (3 digits) is used to classify disease groups with common characteristics. Drug codes are also composed of 5 or more digits and use a 3-unit classification (3 digits) to categorize groups of diseases with common characteristics to reduce the size of the dictionary. (Table 3) Special tokens are tokens required for model training, such as sequence length and exception tokens, and there are five types of them as follows. [pad], [cls], [unk], [sep], [mask] Added hypertension history tokens ([HTN_O]([7]), [HTN_X]([8])) and dyslipidemia history tokens ([LDL_O]([9]), [LDL_X]([10])) for diabetes-based complication models and diabetes

연세대학교
YONSEI UNIVERSITY

history tokens ([DM_O]([5]) for hypertension-based complication models, [DM_X]([6])), and dyslipidemia history tokens for dyslipidemia-based complication models, and diabetes history tokens and hypertension tokens for hypertension-based complication models. Since sensitive diseases are sometimes represented by only one letter of the alphabet rather than the exact diagnosis code, we used the temporary diagnosis code: as a diagnosis code to consider this. (e.g. F: Mental illness) Therefore, the size of the dictionary is 2113 (medical code) + 9533 (pharmaceutical code) + 5 (special) + 26 (temporary medical code) = 11677.

Table 3. Dictionary examples

Variables	Definition/Meaning	Example
DM (HTN, HL, STROKE, CHD)	Whether or not the condition occurred (corresponding statement) Assume '1' for all after the first diagnosis	1
EVENT	Whether the condition occurred (all time periods) Assume '1' for all statements for disease developers	1
DM_first (HTN_firt, HL_first)	If the diagnosis condition includes [outpatient visits≥2/year], the date of the first treatment that meets the diagnosis condition.	20150604
MAIN_SICK_F3	First 3 digits of the main diagnosis code	I10
SUB_SICK_F3	First 3 digits sub diagnosis code	E78

연세대학교
YONSEI UNIVERSITY

5) Input Data Preprocessing - The model used in this study is a Transformer-based model, which, unlike existing RNN-based models, is characterized by computing sequence data at once rather than sequentially. For this purpose, a preprocessing process is required to convert the data divided into statement units into one final input form. The model input data size is 1024, and if the input data size is less than 1024, [PAD] token is added to fit the size, and if the input data size is more than 1024, historical data is truncated to fit the size. The input data variables and formats are as follows (Table 4, Figure 4)

Table 4. Input variables

Input Variables	Description
Gender	Male, Female, Exception
Age	Age as of the most recent statement used as input data.
Medical Diagnosis Code	Main diagnosis code, sub-diagnosis code (3 digits)
Medication Code	All billed medication codes (4 digits)
Start date of care	Calculates and applies the age difference (in years) from the most recent statement and penalizes for historical data.
Hospitalization date	If it is an inpatient statement, it is represented by the statement's admission date. If it is an outpatient statement, represented as a 0
Total visits	Total number of visits in the input data period (number of statements)
Total number of hospitalizations	Number of hospitalizations in the input data period (number of hospitalization statements)

연세대학교
YONSEI UNIVERSITY

MID	MAIN_SICK_F3	SUB_SICK_F3	GNL_CD
100001	I09	E78	1239 l 2625 l 4278 l 6408
100002	H52	H52	2039 l 5300
…	…	…	…

↓

[CLS],['I09'],['E78'],['1239'],['2625'],['4278'],['4599'],['6408'],['H52'],['H52'],['2039'],['5300']…

|

Vocabulary
(size : 11677)

↓

[2],[627],[437],[2382],[3760],[5392],[5661],[7228],[634],[634],[3177],[6328]…

Figure 4. Input variables and format. Shows the preprocessing that converts data that is split into statements into one final input. The age, gender, diagnosis, and drug code per statement are converted into a single sentence format.

4. Modelling and validation

To develop a prediction model among those patients, the transformer-based BERT model was used to develop prediction model in the training group. For each patient ID, there may be about N statements as shown in Figure 5, but the diagnosis and prescription medication that occur in each statement will be different each time, and there will be a difference in the time of the visit, so we represented them as input, age sequence, and applied them to the following model. Since the input of the transformer is all elements at once, we used age as a positional encoding value to distinguish visits. We also used two attentions by applying the weight obtained through multi-head attention to custom attention once more.

Figure 5. BERT architecture. For each patient ID, there may be about N statements. But the diagnosis and prescription medication that occur in each statement will be different each time, and there will be a difference in the time of the visit, so we represented them as input, age sequence, and applied them to the following model

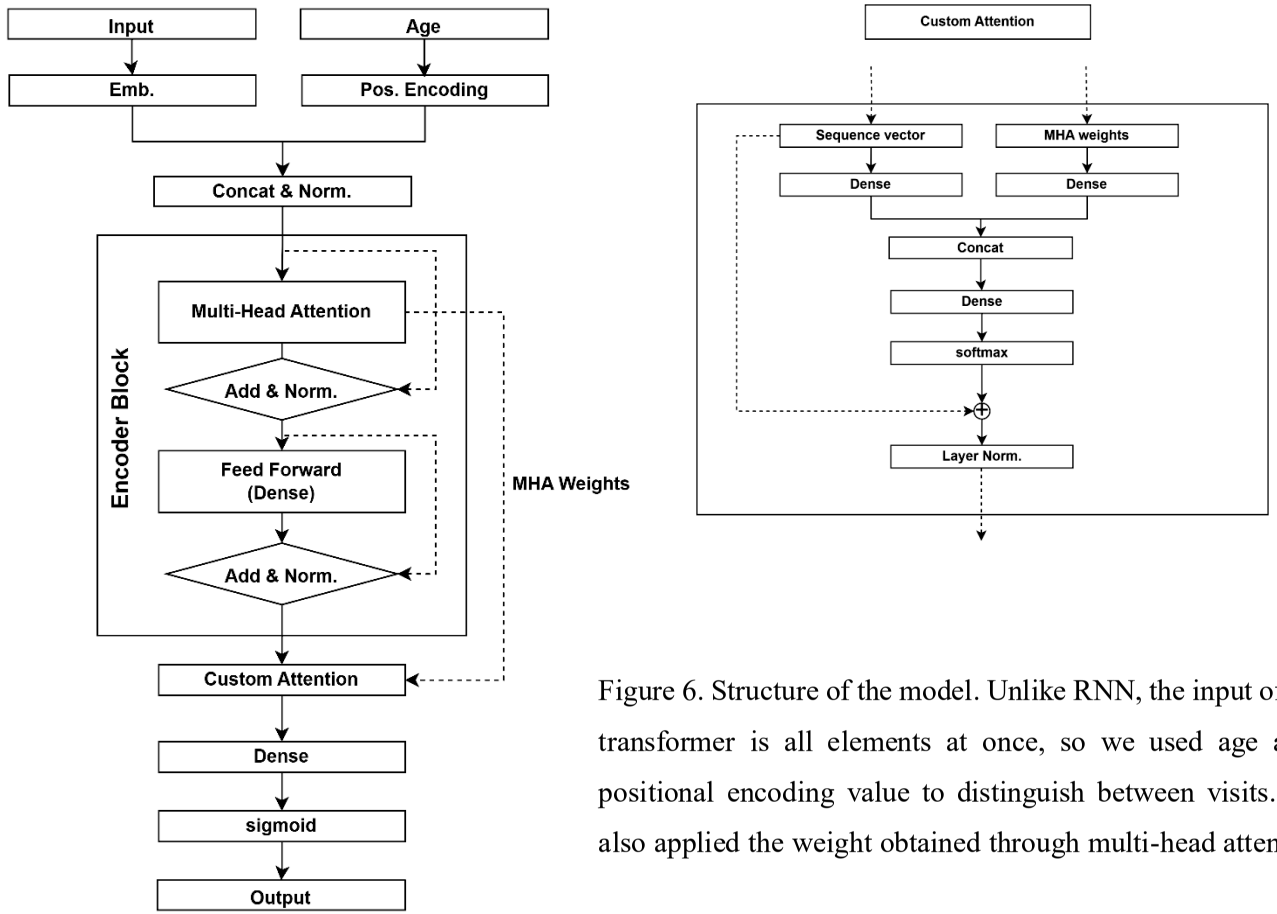

Figure 6. Structure of the model. Unlike RNN, the input of the transformer is all elements at once, so we used age as a positional encoding value to distinguish between visits. We also applied the weight obtained through multi-head attention

연세대학교
YONSEI UNIVERSITY

5. Model Evaluation Method

The evaluation of the disease risk prediction model is measured by the Area Under the Curve (AUC). For the existing multi-classification label, the OvR method (One vs Rest) is used to obtain the AUC, and then the AUC for each class is measured by switching to the binary classification by assuming a specific class as 1 and the rest as 0, and then the AUC for all classes is averaged. However, since the purpose of this study is not to predict the year of disease onset, but to predict the probability of disease onset over 10 years, the AUC value is measured within N years, even though the labels are multi-classified.

연세대학교
YONSEI UNIVERSITY

III. RESULTS

1. Baseline characteristics

The total number of sets used for training and the number of patients with newly diagnosed hypertension, diabetes, hyperlipidemia, and cardiovascular diseases in the sets are as follows Figure 7

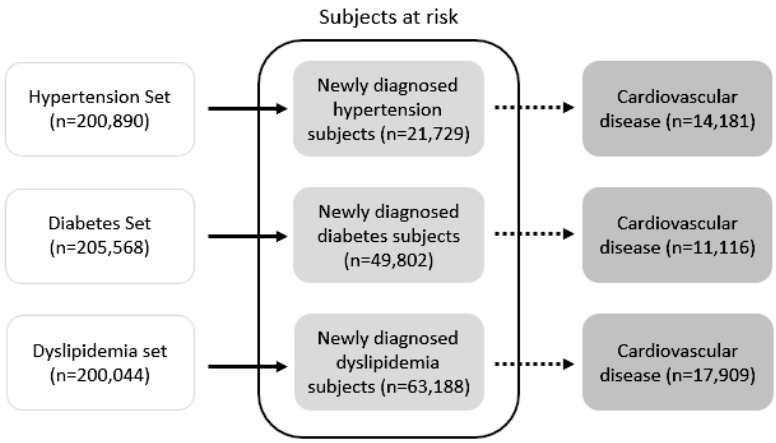

Figure 7. Subjects at risk; newly diagnosed hypertension, diabetes and dyslipidemia. This figure shows the number of new cases of hypertension, diabetes, and dyslipidemia in each of the hypertension, diabetes, and dyslipidemia sets, and the number of cardiovascular diseases that occurred secondary to the diagnosis of each condition.

For each disease, we divided the training set, validation set, and test set into a training set, modeled through the training set, and selected the model with the lowest loss value in the validation set. This model was applied to the test set that was not included in the training set and validation set to obtain the AUC value. The number of positive and negative data used for each disease is in Table 5.

연세대학교
YONSEI UNIVERSITY

Table 5. The number of training, validation, and test sets after augmentation by diseases

	Hypertension		Diabetes		Dyslipidemia	
	Positive	Negative	Positive	Negative	Positive	Negative
Training set	18167	136392	8607	78484	16423	182783
Validation set	2600	19470	1232	11225	2356	26085
Test set	5194	38956	2463	22412	4693	52210
Total	25961	194818	12302	112121	23472	261078

We compared the performance of LSTM, GRU, and BERT, which are models for time series data, on diabetes set and the results are shown in Table 6 and Figure 8. BERT model shows superior predictive performance in prediction of cardiovascular diseases with respect to LSTM and GRU.

Table 6. Model performance in the prediction of cardiovascular diseases according to model

	AUC	Accuracy
BERT	0.905	0.847
GRU	0.835	0.857
LSTM	0.819	0.865

연세대학교
YONSEI UNIVERSITY

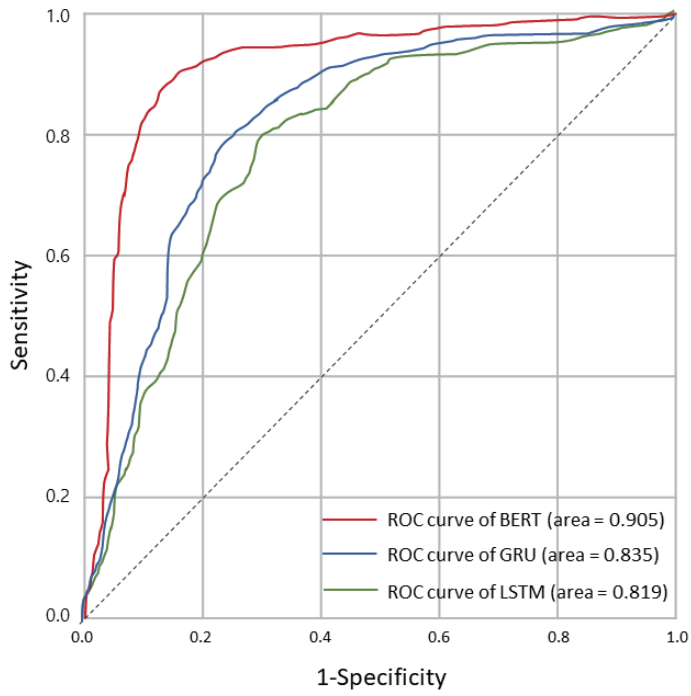

Figure 8. The Receiver operating characteristic curve and area under curve according to model. Model performance in prediction of cardiovascular disease. Compared the performance of LSTM, GRU and BERT. Red line means ROC curve of BERT (area = 0.905), Blue line means ROC curve of GRU (area = 0.835), Green line means ROC curve of LSTM (area = 0.819)

The mean of the cohort was 54 years old and the proportion of men and women was similar across the three datasets. For follow-up time, the positive data is the average of the time between the diagnosis of hypertension, diabetes, and dyslipidemia and the diagnosis of a cardiovascular event, while the negative data is the total time between the diagnosis of hypertension, diabetes, and dyslipidemia and follow-up because no cardiovascular event occurred. The number of outpatient visits appears to be higher in the positive data (the

연세대학교
YONSEI UNIVERSITY

group that developed cardiovascular disease), but when comparing the density of the number of visits divided by the follow-up period, it tends to be higher in the positive data. (Table 7)

Table 7. Baseline characteristics of the group that did (positive) or did not (negative) develop a cardiovascular disease in patients at risk (newly diagnosed diseases; hypertension, diabetes, dyslipidemia)

	Hypertension		Diabetes		Dyslipidemia	
	Positive	Negative	Positive	Negative	Positive	Negative
Gender (Male) [1]	15064 (58%)	110187 (57%)	7602 (62%)	66582 (59%)	13060 (56%)	136102 (52%)
Age [2]	61 ± 10	55 ± 11	61 ± 10	55 ± 11	60 ± 10	52 ± 12
Follow-up duration (year) [3]	3.7 [1.2, 5.8]	5.9 [2.9, 9.6]	3.5 [1.2, 5.4]	6.0 [2.3, 7.7]	3.4 [1.1, 5.3]	5.3 [2.7, 7.9]
Hospitalization duration (day) [3]	1.2 [0, 1]	1.6 [0, 1]	1.3 [0, 1]	1.5 [0, 1]	1.2 [0, 1]	1.3 [0, 1]
OPD visit day (n) [3]	101.5 [21, 137]	131.2 [41, 175]	103.4 [24, 136]	117.2 [36, 154]	93.4 [20, 125]	110.0 [33, 146]
Density [3]	36.1 [15.0, 39.0]	23.9 [12.3, 27.9]	36.2 [16.0, 40.5]	26.1 [13.2, 29.4]	37.1 [14.9, 39.4]	22.5 [11.0, 26.4]

Hospitalization density [3]	0.8 [0, 0.3]	0.4 [0, 0.3]	0.5 [0, 0.4]	0.4 [0, 0.3]	0.7 [0, 0.4]	0.3 [0, 0.3]
OPD visit density [3]	35.3 [14.6, 38.3]	27.2 [12.2, 27.5]	35.7 [15.7, 39.9]	24.7 [13.0, 29.0]	36.4 [14.6, 38.6]	22.2 [10.9, 26.0]
Tertiary general hospital [4]	139176 (5.2%)	1331225 (5.1%)	84203 (6.5%)	850438 (6.4)	138412 (6.2)	1677928 (5.8%)
General hospital [4]	269389 (10.1%)	2260074 (8.7%)	152186 (11.8%)	1423522 (10.7%)	240888 (10.9%)	2728037 (9.4%)
Hospital, Clinic [4]	2136338 (80.1%)	21450945 (83.0%)	1011358 (78.4%)	10743372 (80.7%)	1780002 (80.2%)	24122805 (83.0%)

[1]Number of males and percentage

[2]mean ± SD

[3]mean [Q1, Q3]

[4]Number of statements and percentage

연세대학교
YONSEI UNIVERSITY

2. The performance of the BERT

The performance of the BERT model is shown in the following Table 4. It predicts the occurrence of secondary cardiovascular diseases during the follow-up period after batch inputting and training with age, gender, and prescription drugs for major injuries. The model performed well with an AUC of over 0.9. The performance of the model for predicting cardiovascular diseases is shown in the following Table 5 by adding the past history of each disease during the washout period (e.g., for hypertension, if the patient had dyslipidemia or diabetes before the diagnosis of hypertension) as a past history token

Table 8. Performance metrics for BERT model for predicting cardiovascular diseases

	AUROC	F1 score	Sensitivity	Specificity	Accuracy	Precision	Recall
Hypertension	0.907	0.850	0.803	0.857	0.844	0.671	0.803
Diabetes	0.905	0.840	0.804	0.648	0.847	0.603	0.804
Dyslipidemia	0.904	0.840	0.802	0.851	0.844	0.550	0.802

Table 9. Performance metrics for BERT model for predicting cardiovascular disease using past history token

	AUROC	F1 score	Sensitivity	Specificity	Accuracy	Precision	Recall
Hypertension	0.979	0.851	0.964	0.891	0.910	0.762	0.964
Diabetes	0.978	0.823	0.952	0.895	0.908	0.725	0.952
Dyslipidemia	0.978	0.794	0.949	0.900	0.909	0.682	0.949

연세대학교
YONSEI UNIVERSITY

3. Self-attention score

Figure 8 is an example of self-attention predicting the development of a secondary condition in a patient with newly diagnosed hypertension, showing the attention score from the input data

The top 20 Attention Scores for each disease according to the convergence matrix are shown in the Table 10-15. Common to all three conditions are hypertension, diabetes, and dyslipidemia and their associated medications. In addition, we observed diagnoses or drug codes related to upper respiratory infection or arthritis, or gastrointestinal disease, which are relatively common in the elderly

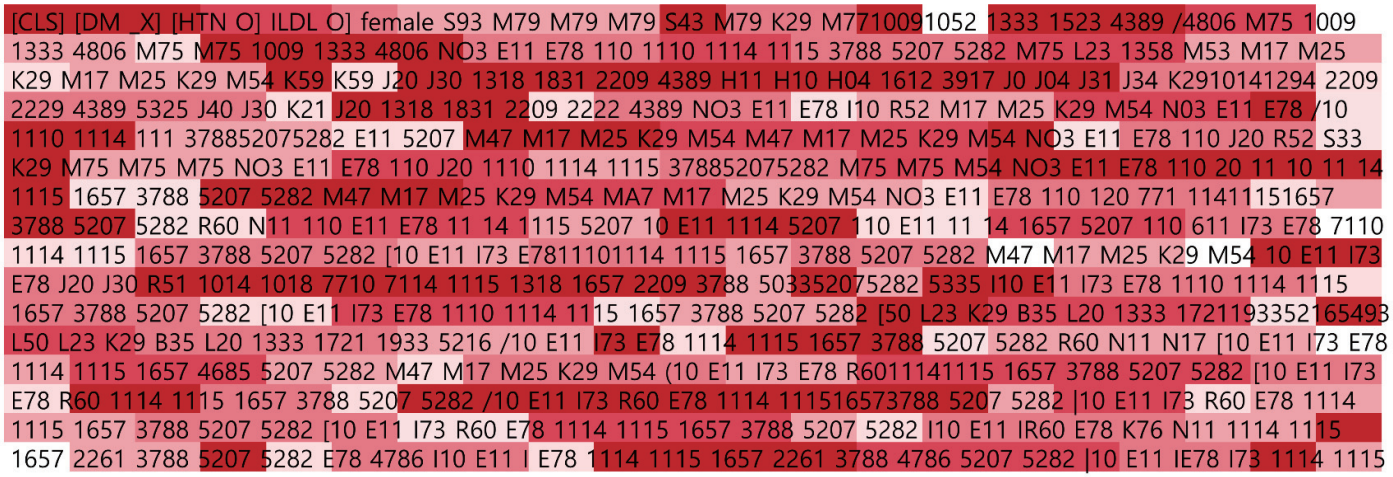

Figure 9. Self-attention example. The self attention example shows the attention score of a patient's input data.

[CLS] is a classification token, which means to classify the entire sequence. [DM X] means there is no history of diabetes, and it is data of a female patient with a history of hypertension and dyslipidemia. Each statement is in chronological order with diagnosis and medication codes, and the correlation between many diagnosis and medication codes is identified, and complications (cardiovascular) are predicted based on the relationship between the codes, and the attention score is displayed in red for high and light color for relatively low. This is an example of predicting cardiovascular diseases based on the relationship between codes with a high attention score in red and a relatively low attention score in light color.

Table 10. The top 20 factors by attention score according to the confusion matrix value – hypertension set

True Positive		False Positive		True Negative		False Negative	
Input	Counts	Input	Counts	Input	Counts	Input	Counts
I10	206966	I10	69710	I10	327248	I10	8443
				E11	226896	E11	4049
E11	76347	E11	24850	1110	168220		
K29	50605	4389	13102	4389	156743	E78	1237
1110	48966	1110	10513			[HTN_X]	1184
4389	41532	K29	9911	2680	90696	4389	1136
1333	24379	J30	9113	2718	85597	J30	1014
E76	22932	M54	7800	E78	83232	K29	996
1115	22625	H04	7487	1333	63083	1915	862
2680	22388	1115	7238	1835	59278	1115	823
M54	21998	2680	7030	K29	54107	2718	746
I20	21647	2718	6935	H04	53290	M79	655
J30	20985	M48	6612	M17	52086	1333	645
H04	20130	M79	6092	1790	49884	J20	645
1915	19256	1915	6061	H25	48119	H04	642
N40	19022	N40	5815	I20	47666	F	604
2718	18356	1333	5502	M48	46278	M54	598
M48	18092	E78	5191	1708	39908	1110	567
M79	17298	J06	5037	J30	39065	2680	564
1835	17171	L23	5008	1115	38412	M81	528

Table 11. The top 20 diagnoses or medications by attention score according to the confusion matrix value – hypertension set

True Positive		False Positive		True Negative		False Negative	
Input	Counts	Input	Counts	Input	Counts	Input	Counts
Hypertension	206966	Hypertension	69710	Hypertension	327248	Hypertension	8443
				Diabetes	226896	Diabetes	4049
Diabetes	76347	Diabetes	24850	Aspirin	168220		
Gastritis	50605	Streptokinase	13102	Streptokinase	156743	Dyslipidemia	1237
Aspirin	48966	Aspirin	10513			[HTN_X]	1184
Streptokinase	41532	Gastritis	9911	Methylephedrine	90696	Streptokinase	1136
Cimetidine	24379	Rhinitis	9113	Ranitidine	85597	Rhinitis	1014
Disorders of glycosaminoglycan metabolism	22932	Radiculopathy	7800	Dyslipidemia	83232	Gastritis	996
Atorvastatin	22625	Disorders of lacrimal system	7487	Cimetidine	63083	Metformin	862
Methylephedrine	22388	Atorvastatin	7238	Levosulpiride	59278	Atorvastatin	823
Radiculopathy	21998	Methylephedrine	7030	Gastritis	54107	Ranitidine	746
Angina pectoris	21647	Ranitidine	6935	Disorders of lacrimal system	53290	Rheumatism	655
Rhinitis	20985	Spondylopathies	6612	Gonarthrosis	52086	Cimetidine	645
Disorders of lacrimal system	20130	Rheumatism	6092	Itopride	49884	Acute bronchitis	645
Metformin	19256	Metformin	6061	Senile cataract	48119	Disorders of lacrimal system	642

Hyperplasia of prostate	19022	Hyperplasia of prostate	5815	Angina pectoris	47666	F	604
Ranitidine	18356	Cimetidine	5502	Spondylopathies	46278	Radiculopathy	598
Spondylopathies	18092	Dyslipidemia	5191	Hydrochlorothiazide	39908	Aspirin	567
Rheumatism	17298	Acute URI	5037	Rhinitis	39065	Methylephedrine	564
Levosulpiride	17171	Allergic contact dermatitis	5008	Atorvastatin	38412	Osteoporosis	528

Table 12. The top 20 factors by attention score according to the confusion matrix value – diabetes set

True Positive		False Positive		True Negative		False Negative	
Input	Counts	Input	Counts	Input	Counts	Input	Counts
I10	64085	I10	26967	I10	206515	I10	3697
E11	36755	E11	18979	E11	136642	E11	3516
				E78	99075	E78	2389
4389	24936	E78	9166			K29	1207
1110	23235	4389	8008	1915	57372		
E78	15701	K29	7318	K29	55605	1110	784
I20	14446	1110	6077	4389	53230	1115	661
K29	13739	2680	5133	1657	49272	1915	581
2680	13664	M54	4785	1110	45368	4389	567
1657	12588	2718	4637	1115	37071	N40	536
1333	12238	N40	4293	2680	27502	[DM_X]	493
2718	12147	H04	3586	2718	26589	2718	470
N40	10826	M79	3457	2229	25863	1657	435
G63	10714	1333	3379	N40	24686	M54	421
M54	9904	J20	3298	M54	24074	J20	394
H36	8995	1657	3280	H36	21402	H04	382
H04	7830	M48	3119	1333	21282	2680	333
2344	7756	I20	3063	M48	19667	L23	331
1835	7352	1115	2834	K76	18887	4786	313
2228	7231	H36	2831	J20	18451	G63	306

Table 13. The top 20 diagnoses or medications by attention score according to the confusion matrix value – diabetes set

True Positive		False Positive		True Negative		False Negative	
Input	Counts	Input	Counts	Input	Counts	Input	Counts
Hypertension	64085	Hypertension	26967	Hypertension	206515	Hypertension	3697
Diabetes	36755	Diabetes	18979	Diabetes	136642	Diabetes	3516
				Dyslipidemia	99075	Dyslipidemia	2389
Streptokinase	24936	Dyslipidemia	9166			Gastritis	1207
aspirin	23235	Streptokinase	8008	Metformin	57372		
Dyslipidemia	15701	Gastritis	7318	Gastritis	55605	Aspirin	784
Angina pectoris	14446	Aspirin	6077	Streptokinase	53230	Atorvastatin	661
Gastritis	13739	Methylephedrine	5133	Glimepiride	49272	Metformin	581
Methylephedrine	13664	Radiculopathy	4785	Aspirin	45368	Streptokinase	567
Glimepiride	12588	Ranitidine	4637	Atorvastatin	37071	Hyperplasia of prostate	536
Cimetidine	12238	Hyperplasia of prostate	4293	Methylephedrine	27502	[DM_X]	493
Ranitidine	12147	Disorders of lacrimal system	3586	Ranitidine	26589	Ranitidine	470
Hyperplasia of prostate	10826	Rheumatism	3457	rebamipide	25863	Glimepiride	435
Polyneuropathy	10714	Cimetidine	3379	Hyperplasia of prostate	24686	Radiculopathy	421

Radiculopathy	9904	Acute bronchitis	3298	Radiculopathy	24074	Acute bronchitis	394
Diabetic retinopathy	8995	Glimepiride	3280	Diabetic retinopathy	21402	Disorders of lacrimal system	382
Disorders of lacrimal system	7830	Spondylopathies	3119	Cimetidine	21282	Methylephedrine	333
Talniflumate	7756	Angina pectoris	3063	Spondylopathies	19667	Allergic contact dermatitis	331
Levosulpiride	7352	Atorvastatin	2834	Liver disease - fatty liver	18887	Omega-3	313
Ranitidine	7231	Diabetic retinopathy	2831	Acute bronchitis	18451	Polyneuropathy	306

Table 14. The top 20 factors by attention score according to the confusion matrix value – dyslipidemia set

True Positive		False Positive		True Negative		False Negative	
Input	Counts	Input	Counts	Input	Counts	Input	Counts
I10	118578	K29	63264	I10	495560	I10	8101
K29	111739	I10	59166	K29	483512	K29	7954
4389	85931						
		4389	37771	E11	171805	4389	2771
1115	52058	1115	24771	4389	170425	J30	2333
1110	48116	2718	22142	1115	163355	1115	2298
I20	47852	J30	17333	J30	157675	2718	2159
2718	42427	M54	15340	M54	143614	E11	2108
2680	33758	1110	15090	2680	93814	M54	1740
1333	33366	2680	14776	M79	93714	2680	1342
M54	27067	E11	13057	2718	92625	E78	1007
J30	26940	1333	12144	E78	86164	1110	1001
2344	24023	I20	9910	J20	85134	H04	986
E11	23375	2228	7577	1110	84641	1333	978
2228	21867	J06	7441	H04	82716	M79	894
2292	16260	H04	7228	L23	61646	J06	855
J06	14873	2344	7059	J06	58036	J20	800
I25	14225	1831	6484	1333	54563	L23	793
1831	13351	M79	6466	M48	53990	1831	699
M48	12408	E78	6394	I20	50708	N40	699

Table 15. The top 20 diagnoses or medications by attention score according to the confusion matrix value – dyslipidemia set

True Positive		False Positive		True Negative		False Negative	
Input	Counts	Input	Counts	Input	Counts	Input	Counts
Hypertension	118578	Gastritis	63264	Hypertension	495560	Hypertension	8101
Gastritis	111739	Hypertension	59166	Gastritis	483512	Gastritis	7954
Streptokinase	85931						
		Streptokinase	37771	Diabetes	171805	Streptokinase	2771
Atorvastatin	52058	Atorvastatin	24771	Streptokinase	170425	Rhinitis	2333
Aspirin	48116	Ranitidine	22142	Atorvastatin	163355	Atorvastatin	2298
Angina pectoris	47852	Rhinitis	17333	Rhinitis	157675	Ranitidine	2159
Ranitidine	42427	Radiculopathy	15340	Radiculopathy	143614	Diabetes	2108
Methylephedrine	33758	Aspirin	15090	Methylephedrine	93814	Radiculopathy	1740
Cimetidine	33366	Methylephedrine	14776	Rheumatism	93714	Methylephedrine	1342
Radiculopathy	27067	Diabetes	13057	Ranitidine	92625	Dyslipidemia	1007
Rhinitis	26940	Cimetidine	12144	Dyslipidemia	86164	Aspirin	1001
Talniflumate	24023	Angina pectoris	9910	Acute bronchitis	85134	Disorders of lacrimal system	986
Diabetes	23375	Ranitidine	7577	Aspirin	84641	Cimetidine	978
Ranitidine	21867	Acute URI	7441	Disorders of lacrimal system	82716	Rheumatism	894
Sodium hyaluronate	16260	Disorders of lacrimal system	7228	Allergic contact dermatitis	61646	Acute URI	855
Acute URI	14873	Talniflumate	7059	Acute URI	58036	Acute bronchitis	800

Atherosclerotic heart disease	14225	Levodropropizine	6484	Cimetidine	54563	Allergic contact dermatitis	793
Levodropropizine	13351	Rheumatism	6466	Spondylopathies	53990	Levodropropizine	699
Spondylopathies	12408	Dyslipidemia	6394	Angina pectoris	50708	Hyperplasia of prostate	699

연세대학교
YONSEI UNIVERSITY

IV. DISCUSSION

In this paper, we introduced a new deep neural network model called BERT to predict the occurrence of major cardiovascular events in patients with newly diagnosed hypertension, diabetes, and dyslipidemia, known as cardiovascular risk diseases, using healthcare big data.

In machine learning, it is well known that learning from more data can improve prediction accuracy, but despite the limited availability of the original data used in this study, augmenting it with data that can be learned through various methods showed superior predictive power compared to other known methods.

Various models have been proposed to predict risk for cardiovascular disease (CVD) in the context of primary prevention.[32-37] Nevertheless, the effectiveness of CVD prediction models that rely on risk factors or statistics is not universally accepted.[38,39] This is because they tend to over- or under-predict risk based on race or socioeconomic status. Therefore, there are ongoing efforts to improve their predictive ability, including the discovery of new biomarkers.

However, the studies on the usefulness of these prediction models and markers did not include Koreans, and there are various opinions on their predictive power in Korean people. Therefore, this model is expected to have a better fit compared to other models because it is based on Korean people.

A recent study showed that traditional risk factors derived from traditional cohort studies, such as body mass index, total cholesterol, blood pressure, and glucose levels, did not appear as significant predictors of cardiovascular disease (CVD) in regression models, which is consistent with previous conclusions from scrutinizing a variety of hospital information, indicating that several customary risk factors have declined in importance

연세대학교
YONSEI UNIVERSITY

with respect to the occurrence of CVD.[40]

Given this, it makes sense that deep learning might be better suited to scrutinizing complex, time-varying data obtained from standard clinical procedures. Such data can be quite different from information obtained through prospective controlled clinical trials.

Previous studies have shown that BERT is an appropriate tool for analyzing EHRs.

With the introduction of electronic health records (EHRs) decades ago, the healthcare field has accumulated a significant amount of electronic health data, and this study confirms the usefulness of BERT models in analyzing large and complex health data sets.

The primary goal of the study was to provide the field with a precise model capable of predicting future disease development. In achieving this goal, BERT produced a number of ancillary results, each of which has independent utility and the potential to be pivotal in subsequent research efforts. In more detail, the disease embeddings extracted through BERT provide profound insights into the interconnectedness of various factors. These embeddings go beyond the mere co-occurrence of diseases and delve into the realm of understanding the proximity of diseases based on trajectories across a broad patient population.

Furthermore, these pre-trained disease embeddings can be used as reliable disease vectors that can be easily deployed by future researchers for numerical and algebraic manipulations.

We have also demonstrated that the disease associations produced by BERT's attention mechanism are useful for explaining disease trajectories in patients with multiple diseases, which not only highlights the co-occurrence of diseases, but also explains the impact of certain diseases in the past on the risk of other diseases in the future.

The idea that a patient's healthcare utilization trajectory and subsequent diagnoses and prescriptions can be used to predict future events, even in the absence of information about

연세대학교
YONSEI UNIVERSITY

known risk factors, is groundbreaking.

For each disease, the self-attention score according to the confusion matrix showed that the well-known hypertension, diabetes, and dyslipidemia were the highest ranked diagnosis or drug prescription codes. In addition, upper respiratory tract infections, gastrointestinal related diseases, and degenerative arthritis, which are diseases that increase with age, were also observed at the top of the list, indirectly indicating that "age" is being reemphasized as an important risk factor.

The performance of the model used in this study was found to be comparable to or better than the performance of models used in other studies that have used BERT to predict disease, particularly cardiovascular disease.[41]

Based on this study, it is expected that BERT can be used as a personalized predictive healthcare model to predict cardiovascular events in patients at risk for cardiovascular disease. By presenting the results of this evaluation, it is expected to improve the healthcare utilization behavior of healthcare users to prevent cardiovascular disease morbidity and improve prognosis. The model is also expected to help in the application of precision medicine.

Limitations

Despite the fact that national health insurance claim data has almost all the information of the entire population, due to some limitations, we only received and analyzed the information of a relatively small number of patients through a well-extracted method. However, to overcome this, we adopted positive and negative augmentation methods to increase the amount of data that can be learned, and we tried to increase the predictive value by testing various augmentation methods. The training, and validation test sets in this study

연세대학교
YONSEI UNIVERSITY

were all done only within national health insurance claim dataset, and there is a point that the performance of this model could not be checked in other cohorts. The model was trained with a limited number of data points and time periods to predict cardiovascular disease over a 10-year period, which has limitations in providing probabilities for each year. This study is a model that predicts cardiovascular morbidity only at 10 years. However, due to the relatively well-stratified cohort, it is unlikely that the predicted rate of cardiovascular diseases will meaningfully decrease during follow-up beyond 10 years. It is also expected to be a good predictive model for patients who may migrate to serious cardiovascular diseases in a relatively short period of time. The inability to predict death due to limited information related to death can also be said to be a limitation. Recently, various studies have shown that obesity and smoking are very important risk factors for cardiovascular diseases, but the lack of consideration of such risk factors in this study is a limitation.

연세대학교
YONSEI UNIVERSITY

V. CONCLUSION

This study introduced a machine learning model called BERT to predict the occurrence of major cardiovascular events in patients with newly diagnosed hypertension, diabetes, and dyslipidemia, known as cardiovascular risk diseases, using healthcare big data. It is a prediction model made with a dataset for Koreans, with a prediction accuracy of more than 0.9, and it is expected to be helpful in the application of precision medicine in that it can predict the occurrence of diseases using individual medical insurance claim data.

연세대학교
YONSEI UNIVERSITY

REFERENCE

1 Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, Pandey M, Maliakal G, Van Rosendael AR, Beecy AN, Berman DS. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. European heart journal. 2019 Jun 21;40(24):1975-86.

2 Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, Tse D, Etemadi M, Ye W, Corrado G, Naidich DP. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature medicine. 2019 Jun;25(6):954-61.

3 Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, Peng L, Webster DR. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature biomedical engineering. 2018 Mar;2(3):158-64.

4 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature medicine. 2019 Jan;25(1):44-56.

5 Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. Nature medicine. 2019 Jan;25(1):24-9.

6 Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE journal of biomedical and health informatics. 2017 Oct 27;22(5):1589-604.

7 Rahimian F, Salimi-Khorshidi G, Payberah AH, Tran J, Ayala Solares R, Raimondi F, Nazarzadeh M, Canoy D, Rahimi K. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. PLoS medicine. 2018 Nov 20;15(11):e1002695.

연세대학교
YONSEI UNIVERSITY

8 Liang Z, Zhang G, Huang JX, Hu QV. Deep learning for healthcare decision making with EMRs. In2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2014 Nov 2 (pp. 556-559). IEEE.

9 Tran T, Nguyen TD, Phung D, Venkatesh S. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). Journal of biomedical informatics. 2015 Apr 1;54:96-105.

10 Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports. 2016 May 17;6(1):1-0.

11 Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. $\mathtt {Deepr} $: a convolutional net for medical records. IEEE journal of biomedical and health informatics. 2016 Dec 1;21(1):22-30.

12 Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor ai: Predicting clinical events via recurrent neural networks. InMachine learning for healthcare conference 2016 Dec 10 (pp. 301-318). PMLR.

13 Pham T, Tran T, Phung D, Venkatesh S. Deepcare: A deep dynamic memory model for predictive medicine. InAdvances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part II 20 2016 (pp. 30-41). Springer International Publishing.

14 Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. Advances in neural information processing systems. 2016;29.

15 Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering. 2009 Oct 16;22(10):1345-59.

연세대학교
YONSEI UNIVERSITY

16 Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. 2013;26.

17 Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training.

18 Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. arXiv preprint. arXiv preprint arXiv:1802.05365. 2018.

19 Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems. 2019;32.

20 Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.

21 Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. InInternational conference on machine learning 2020 Nov 21 (pp. 1597-1607). PMLR.

22 Sun C, Myers A, Vondrick C, Murphy K, Schmid C. Videobert: A joint model for video and language representation learning. InProceedings of the IEEE/CVF international conference on computer vision 2019 (pp. 7464-7473).

23 Adhikari A, Ram A, Tang R, Hamilton WL, Lin J. Exploring the limits of simple learners in knowledge distillation for document classification with DocBERT. InProceedings of the 5th Workshop on Representation Learning for NLP 2020 Jul (pp. 72-77).

연세대학교
YONSEI UNIVERSITY

24 Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942. 2019 Sep 26.

25 Pires T, Schlinger E, Garrette D. How multilingual is multilingual BERT?. arXiv preprint arXiv:1906.01502. 2019 Jun 4.

26 Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129. 2019 May 17.

27 Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323. 2019 Apr 6.

28 Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-40.

29 Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676. 2019 Mar 26.

30 Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342. 2019 Apr 10.

31 Kim L, Kim JA, Kim S. A guide for the utilization of health insurance review and assessment service national patient samples. Epidemiology and health. 2014;36.

32 Conroy RM, Pyörälä K, Fitzgerald AE, Sans S, Menotti A, De Backer G, De Bacquer D, Ducimetiere P, Jousilahti P, Keil U, Njølstad I. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. European heart journal. 2003 Jun 1;24(11):987-1003.

연세대학교
YONSEI UNIVERSITY

33 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. Bmj. 2007 Jul 19;335(7611):136.

34 D'Agostino RB, Grundy S, Sullivan LM, Wilson P, CHD Risk Prediction Group. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. Jama. 2001 Jul 11;286(2):180-7.

35 Lloyd-Jones DM, Leip EP, Larson MG, d'Agostino RB, Beiser A, Wilson PW, Wolf PA, Levy D. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. Circulation. 2006 Feb 14;113(6):791-8.

36 Pencina MJ, D'Agostino Sr RB, Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the framingham heart study. Circulation. 2009 Jun 23;119(24):3078-84.

37 Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998 May 12;97(18):1837-47.

38 Ramsay SE, Morris RW, Whincup PH, Papacosta AO, Thomas MC, Wannamethee SG. Prediction of coronary heart disease risk by Framingham and SCORE risk assessments varies by socioeconomic position: results from a study in British men. European Journal of Preventive Cardiology. 2011 Apr 1;18(2):186-93.

39 Tillin T, Hughes AD, Whincup P, Mayet J, Sattar N, McKeigue PM, Chaturvedi N, SABRE Study Group. Ethnicity and prediction of cardiovascular disease: performance of QRISK2 and Framingham scores in a UK tri-ethnic prospective cohort study (SABRE—Southall And Brent REvisited). Heart. 2014 Jan 1;100(1):60-7.

연세대학교
YONSEI UNIVERSITY

40 Kennedy EH, Wiitala WL, Hayward RA, Sussman JB. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Medical care. 2013 Mar;51(3):251.

41 Suneetha AR, Mahalngam T. Fine Tuning Bert Based Approach for Cardiovascular Disease Diagnosis. International Journal of Intelligent Systems and Applications in Engineering. 2023 May 17;11(6s):59-66.

연세대학교
YONSEI UNIVERSITY

ABSTRACT (IN KOREAN)

위험인자를 가진 환자에서 심혈관 질환을 예측하는 머신 러닝 기반 모델 개발: 보건의료 빅데이터를 이용한 연구

<지도교수 장혁재>

연세대학교 대학원 의학과

송 신 정

전 세계적으로 심혈관 질환이 증가하면서 막대한 사회적, 경제적 비용이 발생하고 있다. 이에 따라 정밀의료 분야는 개인 맞춤형 예측과 예방을 통해 치료를 개선하는 것을 목표로 연구가 이뤄지고 있다. 한국에서는 거의 모든 국민을 대상으로 하는 건강보험 청구 데이터를 보유하고 있어 의료 이용 행태에 대한 모든 정보를 제공하고 있다. 건강보험 사용자는 간단한 인증 절차를 통해 자신의 데이터에 접근할 수 있는 장점이 있어 이 데이터를 이용하여 개인 맞춤형 위험 요인을 예측하는 데 사용될 수 있다. 최근 자연어 처리 영역에서 양방향 변환기 표현(BERT) 및 관련 모델이 주목을 받고 있으며, 텍스트 도메인을 위해 개발된 BERT 모델은 구조화된 건강보험 청구 데이터의 분석 및 적용에 적합할 것으로 판단하였다. 따라서 본 연구에서는

연세대학교
YONSEI UNIVERSITY

건강보험 청구 데이터를 BERT 모델을 통해 위험인자를 가진 환자에서 심혈관 질환발생을 예측하는 모델을 만들고자 하였다. 고혈압, 당뇨, 이상지질혈증을 새로 진단받은 환자를 위험도를 가진 환자로 정의하였으며, 각 질환에서 심혈관계 질환으로 발생하는 것을 예측하고자 하였다. 각 질환은 데이터 증강을 통해 7:2:1의 비율로 훈련, 검증, 테스트 세트로 나누었다. 환자의 진단과 처방된 약물은 입력 시퀀스로 포함되었으며, 방문을 구분하기 위해 나이를 위치 인코딩에 사용하였으며 모델의 예측 능력은 곡선 아래 면적(AUC)을 측정하여 평가하였다.

위험도를 가진 인구 (고혈압, 당뇨병, 이상지질혈증을 새로 진단받은)에서 BERT의 AUC area는 각각 97.9%, 97.8%, 97.8%에 달하였다. Self-attention의 가장 높은 순위를 차지한 질환은 고혈압, 당뇨병, 이상지질혈증 및 노년층에서 더 흔한 진단 및 약물 치료인 것으로 나타났다. BERT는 비교적 적은 훈련 데이터 세트에서 진단명과 약물 처방만을 사용하여도 훌륭한 심혈관 질환 예측 능력을 보여주었다. 이 연구는 BERT가 개인화된 예측 의료 모델로, 위험도를 가진 – 새로 진단받은 고혈압, 당뇨, 이상지질혈증 환자에서 심혈관계질환의 발생 예측결과를 보여주며, 이를 기반으로 하여 예후를 향상시킬 의료이용행태의 개선 및 개인 맞춤의료의 기반이 될 수 있을 것으로 기대한다.

핵심되는 말: bert, 머신러닝, 대사성 질환, 심혈관 질환, 고혈압, 당뇨병, 이상지질혈증.